UNIVERSITY COLLEGE LONDON

G6PD structure and activity: Potential for development of novel low-cost assays for field detection of G6PD deficiency for malaria management.

Francesco Carbone

A thesis submitted to University College London
for the degree of Doctor of Philosophy

November 2016

Declaration of Authorship

I, FRANCESCO CARBONE, declare that this thesis titled, 'G6PD structure and activity: Potential for development of novel low-cost assays for field detection of G6PD deficiency for malaria management.' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 7-11-2016

Abstract

Glucose-6-Phosphate Dehydrogenase (G6PD) is a cytoplasmic protein involved in the first step of the pentose phosphate pathway, that is the oxidation of glucose-6-phosphate to 6-phosphogluconolactone with reduction of NADP+. Deficiency in G6PD activity may result in the formation of different pathologies such as severe forms of anaemia and respiratory distress. In individuals with *Plasmodium vivax* malaria infection, a depression in G6PD activity greatly increases the toxicity of the drugs used in malaria treatment. Tests to detect G6PD deficiencies already exist, but are relatively costly, difficult to perform, and therefore unlikely to be used in endemic areas. The overall aim of this project is the structural analysis of G6PD variants to provide information that could be used in the development of low-cost and simple-to-use immunological field tests for G6PD. Ideally it would have been possible to identify regions that are markers to reduced, or normal activity, to be selected as antibody targets. Initially the SAAP family of tools were used to find G6PD variants that are associated with structural effects at a phenotype level. The selected variants were then studied with all-atom Molecular Dynamics (MD) experiments and looked for shared behaviours among mutants. The difficulty of collecting extensive simulation data made the interpretation of the results challenging, and incomplete, so a united-atom force field (UNRES) was used both to improve the sampling and to increase the numbers of mutants studied. The collected data suggested that the reduced activity in the mutants is not the result of complete unfolding, but it is more likely the result of a very local disruptions in the protein structure the effects of which influence the overall stability and function of the enzyme. To understand the mechanisms of action of the mutations better, a network analysis using the software *wordom* was performed. The idea was to outline key residues (hubs) of G6PD and observe if and how the mutations were capable of altering the communication pathways between hubs. If a mutation, instead of damaging the structure of G6PD, alters the interaction between two or more hubs in the network, it could explain the linkage between the mutation and the reduced activity in G6PD. The final attempt to characterise G6PD behaviours better was the performance of metadynamics simulations with a focus on the role played by Proline 172. This residue has a critical role in allowing the correct positioning of both the substrate and the co-enzyme.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Andrew Martin for his invaluable guidance and encouragement throughout the period of my doctoral research and for giving me the opportunity to work in his group. I would particularly like to thank him for his patience during the corrections of this thesis. I can understand the pain and for that I am very grateful.

This thesis was co-funded by UCL and the Foundation for Innovative New Diagnostics (FIND). I would like to thank both organisations for their generous support.

I would like to thank my Thesis committee members, Dr. Mark A. Williams and Dr. Andrew Osborne for their feedback and support. I would like to thanks the "Research Computing" for all the support with the clusters.

A huge thanks to my parents and my sister Marta, for their presence and support during these four years.

I would like to thank all the members of the Martin group for always being kind and helpful. Thanks to Tom, Saba and Nouf for their help throughout my research. In particular, thanks to Tom for all the programming advice received over the years.

Special thanks to Sayoni, Ivana and Su for many amazing times, and to everyone in Room 636.

Finally, I am very grateful to my girlfriend Clara, for the endless patience demonstrated during the moments of grumpiness.

Contents

List of Figures

List of Tables

Abbreviations

AAN	**A**mino **A**cid **N**etworks
CoM	**C**entre **o**f **M**ass
CV	**C**ollective **V**ariables
FES	**F**ree **E**nergy **S**urface
ff	**f**orce **f**ield
fs	**f**emto **s**econd
G6P	**G**lucose-**6**-**P**hosphate
G6PD	**G**lucose-**6**-**P**hosphate-**D**ehydrogenase
MC	**M**onte **C**arlo
MCC	**M**atthews' **C**orrelation **C**oefficient
MD	**M**olecular **D**ynamics
MM	**M**olecular **M**echanics
ms	**m**illi **s**econd
MTS	**M**ultiple **T**ime **S**tep
mtu	**m**olecular **t**ime **u**nits
NADP$^+$	**N**icotinamide **A**denine **D**inucleotide **P**hosphate (oxidated)
NADPH	**N**icotinamide **A**denine **D**inucleotide **P**hosphate (reduced)
NVE	Micro canonical ensemble (**N**umber of atoms, **V**olume, **E**nergy)
NPT	Isothermal–isobaric ensemble (**N**umber of atoms, **P**ressure, **T**emperature)
ns	**n**ano **s**econd
NVT	Canonical ensemble (**N**umber of atoms, **V**olume, **T**emperature)
pbc	**p**eriodic **b**oundary **c**onditions
PCA	**P**rincipal **C**omponent **A**nalysis
PES	**P**otential **E**nergy **S**urface
PD	**P**hatogenic **D**eviation
PMF	**P**otential of **M**ean **F**orce
PPP	**P**entose **P**hosphate **P**athway

ps	**p**ico **s**econd
PSN	**P**rotein **S**tructure **N**etwork
PTMetaD	**P**arallel **T**empering **MetaD**ynamics
RBC	**R**ed **B**lood **C**ell
REMD	**R**eplica **E**xchange **M**olecular **D**ynamics
RMSD	*R*oot *M*ean *S*tandard *D*eviation
RMSF	*R*oot *M*ean *S*tandard *F*luctuation
SAS	**S**olvent **A**ccessible **S**urface
SNP	**S**ingle **N**ucleotide **P**olymorphysm
$\boldsymbol{U_{UB}}$	**U**rey-**B**radley
wt	**w**ild-**t**ype
$\boldsymbol{\mu}$**s**	**m**icro **s**econd
$\boldsymbol{\tau_G}$	frequency of Gaussians deposition

Physical Constants

Boltzmann constant	k_B	=	$1.38 \cdot 10^{-23}$ JK^{-1}
Molecular time unit	1 mtu	=	48.9 fs

Chapter 1

Biological background

Malaria is a human infectious disease, caused by a member of the protist family of *Plasmodium*. Among the five species of *Plasmodium*: *P. ovale curtisi, P. ovale wallikeri, P. malariæ, P. vivax* and *P. falciparum*, the latter is the most virulent and deadly [1, 2]. Malaria is notorious for its ability to alter tissue perfusion by causing adhesion of infected red blood cells (RBCs) to the walls of blood vessels; it also causes the eventual destruction of RBCs, thus compromising oxygen delivery. Mosquitoes (genus *Anopheles*) are the vector agents which transmit parasites from person to person without suffering from the disease themselves. Malaria is prevalent in tropical regions where a particularly humid and hot environment provides ideal conditions for mosquitoes to breed. Currently, the only satisfactory strategy to stop transmission of malaria is the prevention of mosquito bites; once the parasites are inside the host body, other pharmaceutical strategies must be adopted.

1.1 Plasmodium pathogenicity & pathopysiology

While all *Plasmodium* types share the same life cycle (Figure 1.1) and pathogenesis, the outcome of malaria depends on many different factors: parasite factors (drug resistance, antigenic polymorphisms or invasion pathways), host factors (age, genetics and immunity) and geographic factors.

The infection starts when a mosquito injects parasites (*sporozoites*) into the host subcutaneous tissue. Parasites travel to the liver where they settle inside the hepatocytes and start their maturation process. After around 48 hours, each sporozoite develops into thousands of *merozoites*, which invade RBCs once released into the bloodstream. With

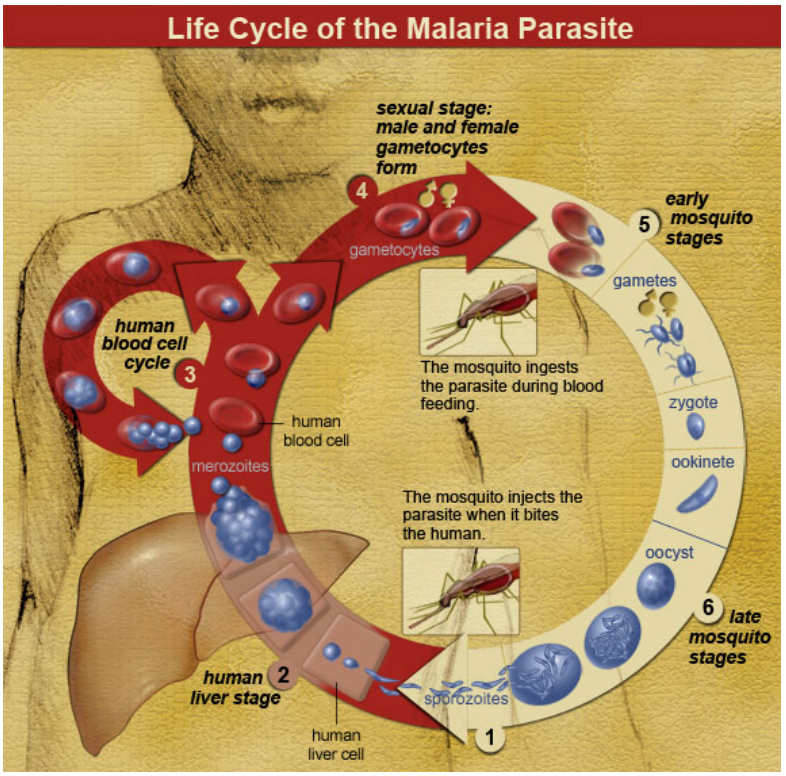


FIGURE 1.1: Malaria parasite life cycle. 1) Following the mosquitoes' bites, sporozoites are released into the bloodstream. 2) Sporozoites travel to the liver and infect liver cells. 3) From the liver, merozoites, the mature asexual form of parasites, invade red blood cells (RBCs) and start to multiply inside RBCs. 4) A portion of parasites converts to gametocytes that are ingested by the mosquito. 5-6) Male and female gametes fuse to form diploid zygotes which develop in thousands of active haploid forms (sporozoites). Taken from the National Institute of Allergy and Infectious Diseases (NIAID).

20 new merozoites being generated per parasite, asexual parasites multiply in RBCs and start the spread of the disease. Only a small portion of the asexual parasites convert to *gametocytes* which are in turn transmitted to the female anopheline mosquito at the next bite. *P. falciparum* is responsible for the largest number of cases of fatal malaria and the belief that *P. vivax* infection was in most cases benign, has led to the majority of studies focusing on *P. falciparum*. However, outside Africa, *P. vivax* accounts for almost half of malaria cases and many studies have shown a strong association between *P. vivax* infection, severe disease and death. *P. vivax* has a preference for *reticulocytes* (14 day old immature blood cells) [3], whereas *P. falciparum* indiscriminately infects all kinds of RBCs. Despite this limitation, inflammatory response during *P. vivax* infection is much greater than that seen in other *Plasmodium* infections [4–6]. *P. vivax* has also shown a lower degree of cytoadherence, causing less organ dysfunction than *P. falciparum*. The advantage of this behaviour may be understood by looking at the deformation property of RBCs; in fact, while *P. falciparum* escapes organs' filtration by inducing cytoadherence, *P. vivax* avoids destruction during passage through endothelial cells by increasing

deformability of infected RBCs [7, 8]. The fundamental difference between *P. vivax* and *P. falciparum* is the ability of *P. vivax* to become active from dormant *hypnozoites* [9, 10]. Frequent activation and patient relapses cause malaria tolerance (higher fever threshold and attenuated symptomatology), but recurring episodes of haemolysis and syserythropoiesis are likely to contribute to a severe form of anaemia. These patients do not have feverish symptoms and so are less likely to seek treatment [4, 11]. Although rare, respiratory distress, acute lung injury and coma may occur [11], but mostly only in the presence of other factors such as concurrent infections, occult mixed plasmodium infections, metabolic changes and micro vascular dysfunction.

1.2 *Plasmodium Vivax* vaccine

Malaria has always been the target of intense campaigns aimed at eradicating parasite transmission. Nevertheless, significant progress is still to be made [12]. The emergence of mosquito strains resistant to insecticides and parasites resistant to anti-malarial drugs are two of the factors making malaria eradication complicated [11, 13–15]. As stated by Arevalo-Herrera *et al*, "*It is important to understand that owing to the complex parasite life cycle and existence of co-infections, it is commonly accepted that, rather than a single specific vaccine, a multi-antigen and multi-species vaccine should be developed*" [16]. Currently three parasite stages have been identified as potential targets for effective anti parasite immune response [16], which are:

1. **The pre-erythrocytic stage:** prevent the entrance of *sporozoites* into hepatocytes and inhibit their development, or develop drugs that would target antigens expected to induce cellular immune response. The only two sporozoite antigens that have been identified are *circumsporozoite protein* (CSP) and *sporozoite surface protein 2*(SSP2/TRAP) [17]. CSP is composed of an immunodominant repeated domain and two highly conserved flanking amino- and carboxy- regions called "Region I" and "Region II-plus" [18–20]. SSP2/TRAP mediates adhesion to host cells and tissue surfaces [21–23].

2. **The asexual erythrocytic stage:** prevent clinical manifestation and severity of the infection, resulting in a reduction of both morbidity and mortality. Efforts are concentrated on antigens present on merozoite surface, better to understand the mechanism of invasion. The most studied antigens are the *Duffy binding*

protein(DBP) [24, 25], the *merozoite surface protein 1*(MSP1) [26–28] and the *Apical membrane antigen 1*(AMA-1) [29–31].

3. **The parasite sexual development:** Another potential course of action is to block the fertilization process and parasite development inside the mosquito [32, 33].

The only drugs capable of clearing the parasites in the liver stage are the *8-aminoquinolines*, such as Primaquine and Tafenoquine (Figure 1.2).

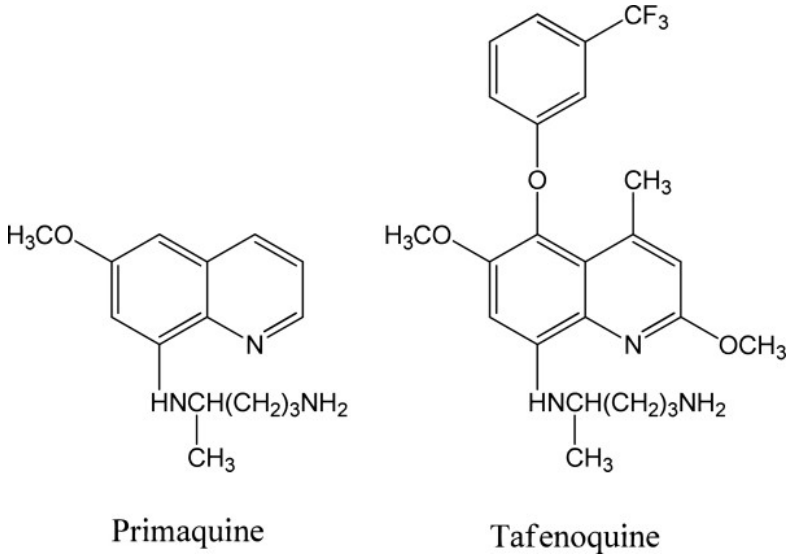


FIGURE 1.2: Chemical structures of primaquine and tafenoquine [34].

1.2.1 Primaquine

Although the mechanism of action of this drug is not yet fully understood, it is thought to interfere with the cellular respiration of the parasites, by increasing the oxygen free radicals and deregulating the electron transport [35]. One of the advantages of Primaquine is that it can be used successfully in both primary (administration before exposure) and terminal (administration after the exposure) prophylaxis. Additionally Primaquine is helpful in preventing the transmission of *P. falciparum*, by eradicating circulating gametocytes. In areas where both *P. vivax* and *P. falciparum* are present, infection with one parasite increases the susceptibility to the other strain, so having an effective drug is a huge boost in malaria management. Treatments with Primaquine show up to 90% efficacy in clearing the hepatic reservoir that causes relapses [36, 37]. Like every drug, the use of Primaquine in malaria treatment has some side effects such

as methemoglobinemia (a disease where hemoglobin is oxidised to methemoglobin, resulting in a reduced ability to release oxygen to tissues), hypersensitivity reactions and gastrointestinal disturbance.

However the most feared hazard is the haemolysis precipitation in G6PD-deficient individuals. G6PD activity screening are available and can determine the level of deficiency, but the size of the exposed population makes this process costly. Besides, with over 300 different allelic mutations for this X-linked recessive disorder, the phenotypes of G6PD deficiency are extremely variable [38, 39], making the complete screening of all possible phenotypes among a given population even more complex. An interesting aspect is that G6PD deficiencies may confer some resistance against *Plasmodium* infections [27]. This can lead to the misconception that most infected individuals are unlikely to present a G6PD deficiency and that the use of Primaquine can be relatively safe. However evolutionary selection has made the number of individuals with G6PD deficiencies increase in endemic areas. It is always recommended that G6PD deficiency is checked before Primaquine treatment [16, 35]. Primaquine is a powerful drug, but apart from the side effects, other factors may influence both the efficacy and toxicity after administration. First of all, to reduce the risk of treatment failure, developing resistance and reducing side-effects, Primaquine is administered together with blood schizontocidal agents, making it difficult to control the final effects. Additionally the dosage varies greatly between individuals: Failure rates decrease with increased dose, but body size, age and ethnic group differences make the generalization of the therapy difficult. Consequently pharmaceutical companies are pushing towards safer drugs that can be used in treatment of the malaria virus. The most promising alternative is Tafenoquine, a compound that has a shorter course of therapy, down from 14 days to 3 days [40–42]. Despite these improvements, the issue of haemolysis within G6PD deficient individuals still remains a risk and there is therefore a need to develop a low-cost test to identify individuals with depressed G6PD activity.

1.3 Glucose-6-Phosphate Dehydrogenase

Glucose-6-phosphate dehydrogenase (G6PD) is a cytoplasmatic protein whose deficiency is the most common human enzymopathy. The enzyme active form is the result of a rapid dimer-tetramer equilibrium, which is affected by ionic strength and pH [43, 44]. The active form of G6PD is involved in the first step of the pentose phosphate pathway (PPP), which is the oxidation of glucose-6-phosphate to 6-phosphogluconolactone with reduction of $NADP^+$ to NADPH (Figure 1.3). This step is the irreversible, controlling step of the PPP, and G6PD activity is allosterically stimulated by concentration of $NADP^+$ and inhibited by concentration of NADPH [45]. The PPP is important in all cells, not only for the production of pentoses (5-carbon sugars), but also for the production of reducing agents in the form of NADPH. NADPH is also essential for protection against oxidative damage by maintaining a high level of reduced glutathione (GSH) that protects the sulphydryl groups in haemoglobin and in red cell membranes from oxidation [46]. When oxidative stress occurs, the level of GSH drops and can be restored by glutathione reductase which requires NADPH to work. G6PD is the only NADPH-producing enzyme that is activated by oxidative stress and, in RBCs, it is also the only source of NADPH. If levels of GSH drop and NADPH is not produced, oxidative damage eventually leads to cell haemolysis.

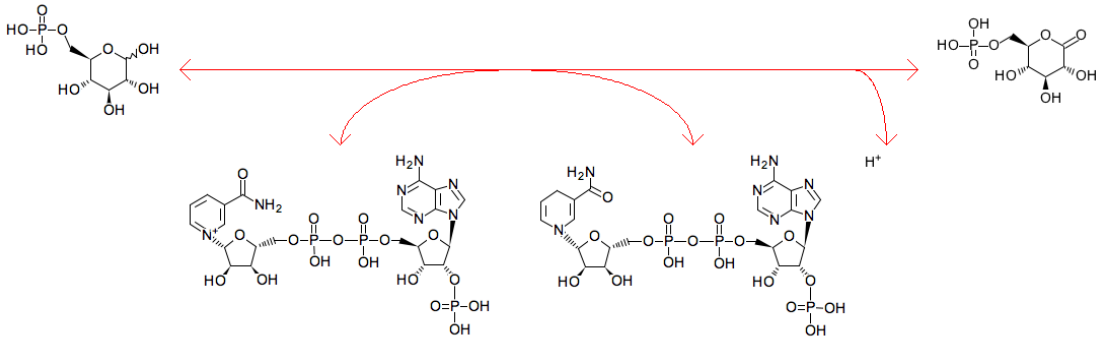


FIGURE 1.3: G6PD reaction diagram, obtained from the KEGG ftp site (ID R00835): Oxidation of the CH-OH group of the D-glucose-6-phosphate to 6-phospho-D-glucono-1,5-lactone, with reduction of $NADP^+$ to NADPH.

1.3.1 G6PD deficiency

The G6PD locus, with over 300 allelic variants, is the most polymorphic enzyme in humans [38, 39]. G6PD deficiency affects over 400 million people worldwide, concentrated mainly in tropical and sub-tropical regions. Comparisons of all variant gene sequences with the normal G6PD gene have identified 186 different mutations, divided between single point mutations (85.4%), multiple mutations (8%), deletions (5.3%) and intron mutations (1%). Since most of the mutations are asymptomatic, the total number of mutations could be much higher. These variants have been grouped into five classes depending on the degree of enzyme activity by the World Health Organization (WHO) [47]:

1. Class I: Severe chronic anaemia;
2. Class II: Severe deficiency (<10% activity), with intermittent hemolysis;
3. Class III: Mild deficiency (10-60% activity), hemolysis with stressors only;
4. Class IV: Non-deficient variant, normal enzyme activity;
5. Class V: Increased enzyme activity.

1.3.1.1 Clinical presentations

Neonatal jaundice and acute and chronic haemolytic anaemia are the principal manifestations of G6PD deficiency [48]. Neonatal jaundice is a disease that occurs in the presence of inappropriate levels of *bilirubin*. Bilirubin is a metabolite of heme catabolism that is bound to albumin for transportation and converted into *glucuronides* in the liver. Here the bilirubin is conjugated with *glucuronic acid* to form *mono* and *di-glucuronides* which are eventually excreted in the bile [49]. In normal neonates, the excess bilirubin is hydrolysed back and reabsorbed into the circulation. If this does not happen, levels of glucuronides grow causing bilirubin encephalopathy which may lead to mental retardation. The role played by G6PD is not yet understood, but it has been proved that G6PD deficiency leads to an increased incidence of neonatal jaundice [49, 50]. Another manifestation is “acute hemolytic anaemia” in which acute episodes of intravascular haemolysis follow the ingestion of certain kinds of food or drugs. As an example, *divicine*, an unstable metabolite of the pyrimidine β-glucoside *vicine*, contained in broad beans, acts as a strong oxidizing agent, triggering the haemolysis effect. The degree of anaemia varies with the nature of the stress and depends on the type of variants. Conversely chronic

nonspherocytic hemolytic anaemia occurs in males with very low level of G6PD activity. In such conditions, acute anaemia is triggered by a wider number of agents and in lower concentrations than in those obtained with other variants. In addition, individuals with G6PD deficiency have a higher probability of developing associated pathological conditions. Sickle cell hemoglobinopathy and RBC thalassemia are frequently associated with G6PD deficiency. The presence of multi-pathology does not alter the course of each independent disease and does not determine synergistic effects [51–55].

1.3.1.2 G6PD variants

A single nucleotide polymorphism (SNP) is defined as a variation in a single nucleotide that occurs in at least 1% of a 'normal' population. Individuals from different geographical areas usually present different sets of polymorphic variants (for a detailed list of the most common variants see Table 1.1). The variant "Mediterranean" (Ser188Phe)[56] is common in the Mediterranean area, the middle East and India; in Africa the main variant is "A^-" (Val68Met + Asn128Asp) [57], while in China several variants are distributed all around the country [58]. It is commonly accepted that the high frequency of polymorphic variants has been generated because of the relative protection against severe malaria that G6PD deficiency provides [59, 60]. Although this protection mechanism is not known, some studies suggest that because of the high levels of oxidative agents, infected erythrocytes are rapidly damaged and destroyed by phagocytosis [61]. An interesting behaviour that was reported in the early biochemical characterization papers is that, while enzyme activity in G6PD-deficiency erythrocytes decreases, the enzyme activity in white cells does not, or only very little [62, 63]. Mutations generating G6PD deficiency are strong enough to affect enzyme activity in RBCs, but not so severe as to decrease efficiency and activity in other somatic cells [64, 65]. The reason behind this behaviour may be that the lack of G6PD in somatic cells is lethal [39, 66].

Name of mutation	Amino acid change(s)	Comments	Distribution
Gaohe	32His→Arg		China,Thailand,Malaysia
Honiara	33Ile→Met, 44Ala→Gly	May be a neutral mutation on a G6PD Union background	Solomon Island
Orissa	454Arg→Cys		Tribal India, Mauritius, Malaysia
Aures	48Ile→Thr		Algeria, Kuwait, Saudi Arabia,UAE, Spain
Metaponto	58Asp→Asn		Italy
A^-	68Val→Met, 126Asn→Asp	Found in people of African origin.	Africa, Spain, Portugal, Middle East, USA, Cuba,Brazil etc
Namoru	70Tyr→His		Tribal India (South)
Ube-Konan	81Arg→Cys	Most common variant in Japan	Japan
Vanua Lava	128Leu→Pro		Malaysia, Indonesia, Vanuatu
Mahidol	163Gly→Ser	Common variant in Thais	China, South East Asia
Santamaria	181Asp→Val, 126Asn→Asp		Spain,Mexico,Costa Rica, Algeria,Sicily
Mediterranean	188Ser→Phe	Most common mutation in many Mediterranean and Middle East countries and in the Indian subcontinent.	India,Malaysia,Italy, Greece, Spain,Portugal,Middle East, Croatia, Brazil etc
Coimbra	198Arg→Cys		Malaysia,Indonesia, Cambodia
Seattle	282Asp→His	Widespread. Only varinant found in a study in Canary Island	Sardinia, Brazil, Mexico, Spain, Portugal, Canary Island, Italy,Greece, Croatia
Montalbano	285Arg→His		Italy
Viangchan	291Val→Met	Most common variant in Thailand and Cambodia	China,Thailand,Malaysia, Cambodia,Laos,Indonesia
Kerala/Kalyan	317Glu→Lys		India
Chatham	335Ala→Thr		India, Indonesia, Italy, Iran,Malaysia, Kuwait, China,Spain, Japan
Chinese-5	342Leu→Phe		China, Malaysia, Thailand and Singapore
Ierapetra	353Pro→Ser		Greece
Cassano	449Gln→His		Italy, Greece
Union	454Arg→Cys	Widespread	Solomon Island, Vanuatu,Croatia, China, Italy, Spain,Mexico, Cambodia, Thailand,Malaysia
Canton	459Arg→Leu	Common amongst Malaysian and Chinese people.	China, South East Asia
Kaiping	463Arg→His	Most common variant in Flores, Indonesia	China, South East Asia, Indonesia

TABLE 1.1: The major polymorphic G6PD variants and their distribution [48].

1.3.2 Screening

WHO report No.366 [47] indicates all the standardised procedures for the study of G6PD deficiencies. The tests to assess G6PD activity can be divided into *genetic* and *phenotypic-biochemical* approaches.

1.3.2.1 DNA-based genotyping approaches

DNA sequencing has successfully identified a great range of G6PD variants, but not all of those are associated with a deficiency, meaning that an additional phenotypic screening is required to determine the clinical relevance of a specific mutation. Studies on Burkinabe populations show how genotyping assays can be successfully used to investigate G6PD deficiency if one or few haplotypes account for more than 90% of the mutations presented in the population [67], and even in such cases, the correlation between genotype and phenotype rarely exceeds 70% [68]. A major drawback is the complexity of these screenings: DNA-based diagnostics can be successfully used to study multiple polymorphisms (e.g. the genetic profile of an entire population), but their costs and complexity make them infeasible for a point-of-care diagnostic.

1.3.2.2 Phenotypic-biochemical approaches

Phenotypic testing is the most informative and frequently used diagnostic methodology for the detection of enzymatic activity levels. Although there are more than 30 testing kits, all of them belong to one of these categories:

- *Direct tests* directly determine G6PD enzymatic activity by spectrophotometry [47]. These methods detect the fluorescence of NADPH under long-wave (365 nm) UV light in complete darkness. Reduction of NADP to NADPH occurs in the presence of G6PD, which means that the rate of NADPH formation is proportional to G6PD activity. Beutler's test, for example, uses a fluorescent spot test, where fluorescence readings are performed at different intervals after incubation, and the samples are classified in different groups based on the intensity of fluorescence.
- *Indirect tests* indirectly detect G6PD activity by analysing substrates that may be associated with G6PD. One of the widest-used tests, the methaemoglobin reduction test (MRT)[69], measures the rate of NADPH-dependent methaemoglobin

reduction in the presence of an appropriate redox agent. The calculated rate then correlates with G6PD activity.

- *Cytochemical assays* assess G6PD activity by labelling each individual erythrocyte. Once the water-soluble colourless tetranitro blue tetrazolium is reduced via the electron carrier 1-methoxyphenazine methosulfate by NADPH, dark-purple granules are present in erythrocytes that contain G6PD activity, whereas G6PD-deficient erythrocytes remain unstained. Cytochemical assays have the advantage of working with individual cells (i.e. only a small sample is required), but of all the listed approaches, they are the most expensive and difficult to perform.

Although all of these methods have proven to be effective in different conditions, neither genetic nor phenotypic methodologies are able of capturing the full picture of G6PD deficiency. Several factors work against the reliability of each of the methods; first of all, the difficult environmental conditions of the endemic countries reduce the accuracy and reliability of all of the G6PD tests. Furthermore the cost of the tests, that has to be added to the cost of the drugs, increases the economic burden on endemic areas. All these problems must be solved in order to win the battle against malaria and developing a new cheap test is an urgent priority.

1.3.3 G6PD Structure Description

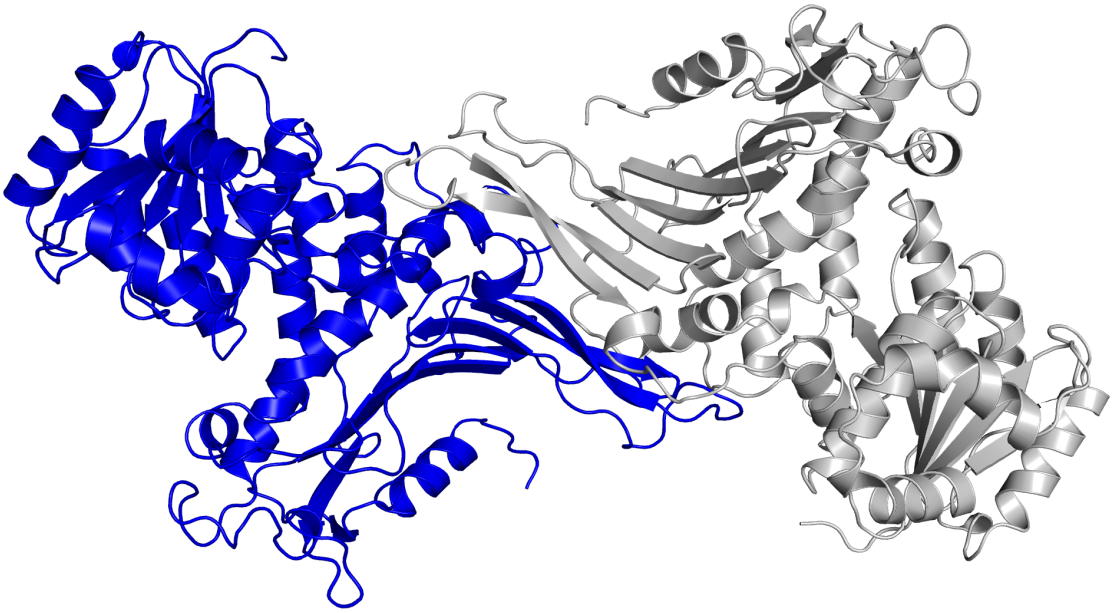


FIGURE 1.4: The G6PD dimer (PDB code 2bhl) with the two monomers coloured in blue and grey.

In humans, G6PD is a 59 kDa protein encoded in the telomere region of the X chromosome (Xq28) [70]. The dimer is composed of two identical 514 amino acid subunits (Figure 1.4), each containing a single active site which binds the glucose-6-phosphate (G6P) and two $NADP^+$ binding sites, one which binds the co-enzyme and the other which has structural importance (Figure 1.5 and Figure 1.6). The active form exists in both dimeric and tetrameric form, with the monomer being inactive [48, 71]. The interconversion between these forms is controlled by the pH and by G6P and NADPH levels. Wrigley *et al.* [72] describe how reducing the pH from 8 to 5.8, causes the proportion of dimer G6PD to drop from 70% to 6%, while, a rise in the proportion of the tetramer from 0 to 90 % is recorded, suggesting that at low pH (below pH 6) the equilibrium is shifted towards the tetramer, while at high pH (above pH 8) most of the enzyme is in the dimeric form [44]. At physiological pH, both forms exist, with the dimer being the more favourable of the two forms. In an abundance of $NADP^+$ and G6P, G6PD activity is required and $NADP^+$ can bind the structural $NADP^+$ binding site, inducing dimer formation by strengthening the interactions between monomers.

There are a total of 13 experimentally determined structures of the G6PD enzyme: 9 from *Leuconostoc mesenteroides*, 1 from *Trypanosoma cruzi* and 3 from *human* (Table 1.2). The first human structure to be crystallised (PDB code 1qki) was the tetramer of the Canton variant (R465L) [73]. This mutant is one of the most common Chinese variants and individuals with this mutation exhibit a class II phenotype [58, 74]. This

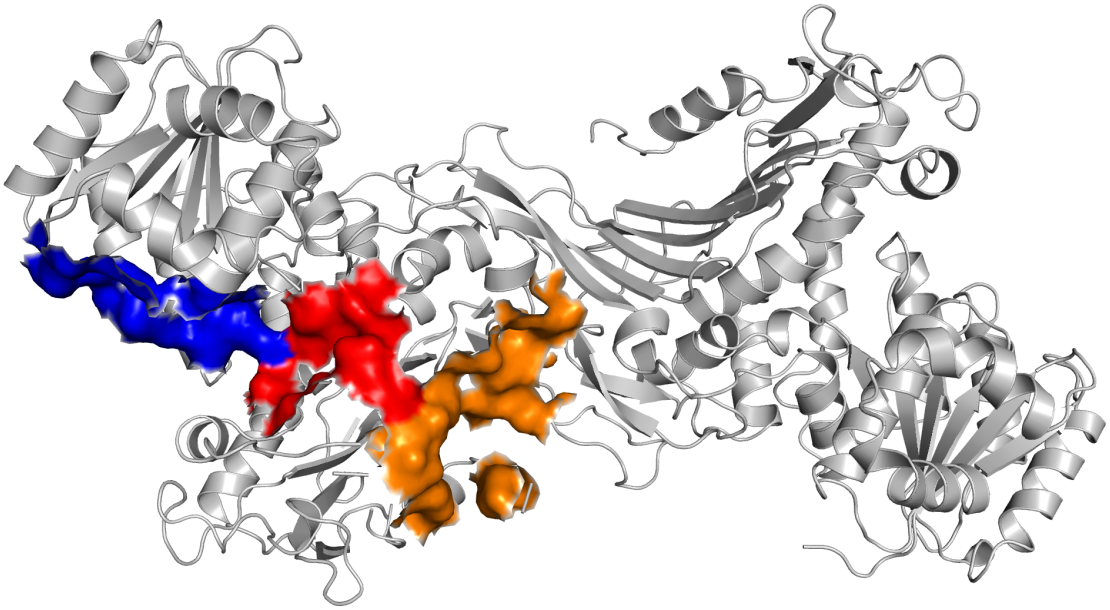


FIGURE 1.5: The G6PD dimer (PDB code 2bhl) with the residues involved in binding represented as a surface. The G6P binding site is coloured in red, while the co-enzyme and the structural $NADP^+$ binding sites are in blue and orange respectively.

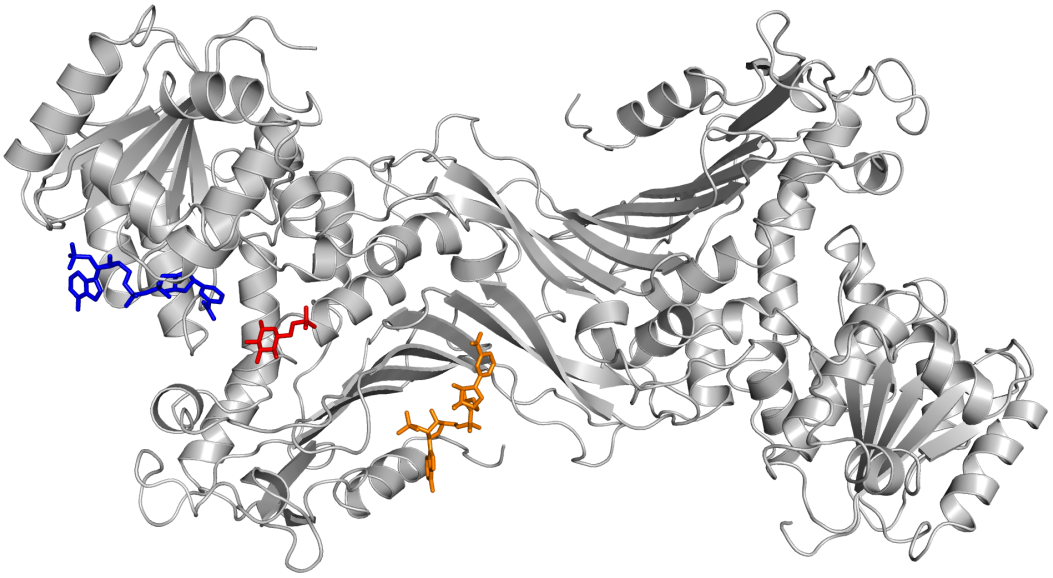


FIGURE 1.6: The G6PD dimer (PDB code 2bhl) bound with the substrates. The G6P is coloured in red, the co-enzyme $NADP^+$ is in blue and the structural $NADP^+$ is in orange.

structure is a dimer of dimers (Figure 1.7), has a resolution of 3 Å, and is in complex with the structural $NADP^+$ that binds the area between the β sheet and the C-terminus (see Section 1.3.3.3). The monomer (PDB code 2bh9) and the dimer (PDB code 2bhl) are both a non-natural variant (ΔG6PD) obtained by the deletion of the 25 N-terminal residues of the wild-type G6PD. The deletion does not influence the enzyme stability

or its activity, and was done to increase the quality of the crystals obtained [73]. The human structures are very similar to that of *L. mesenteroides*, with both substrate and co-enzyme binding presenting only minor differences as a result of sequence changes. The major difference is the fold at the C-terminus of *L. mesenteroides* which has a total absence of the structural $NADP^+$ site. This site is found in rat, mouse, wallaroo, fish (*fugu rubripes*) and fruit flies, but it is absent in all prokaryotic G6PD. It is likely that prokaryotic enzymes have a higher number of replacements, causing the C-terminal tail to shorten making it impossible for the additional $NADP^+$ to bind [73]. Because of the absence of the structural $NADP^+$ in all the *Leuconostoc mesenteroides* structures and because the deletion in ΔG6PD is comparable with the wild-type, it was decided that the human dimer (PDB code 2bhl) was the best structure to be used for all the simulations.

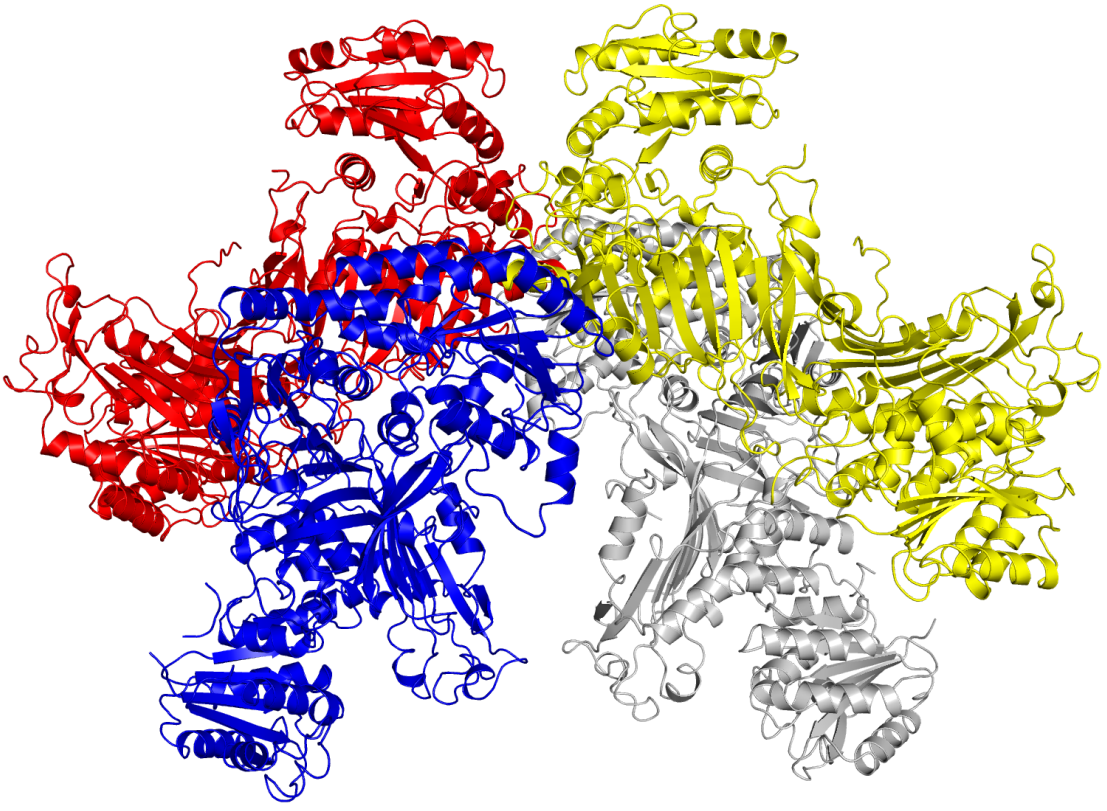


FIGURE 1.7: The G6PD tetramer (PDB code: 1qki) with the four dimers coloured in different colours.

The amino acid sequence of G6PD has been highly conserved through evolution [75, 76] and multiple alignment of 35 known sequences of G6PD indicates a sequence identity greater than 48% among eukaryotic G6PDs sequences. This number drops to a 30% identity when prokaryotic sequences are also considered [77]. In particular three key motifs have been identified in all G6PDs sequences: 198-RIDHYLGKE-206 (Figure 1.8

red), 38-GxxGDLA-44 (Figure 1.8 blue) and 170-EKPxG-174 (Figure 1.8 yellow). The first peptide is the binding and catalytic site for G6P [78, 79], while the other two peptides are involved in $NADP^+$ binding [80, 81].

PDB code	description	res [Å]	R-factor	R-free	ligands
2bh9	monomer (H)	2.5	0.201	0.296	G6P
2bhl	dimer (H)	2.9	0.214	0.261	Co-enzyme and structural $NADP^+$
1qki	tetramer (H)	3	0.247	0.294	structural $NADP^+$
1dpg	dimer (L)	2	0.206	0.257	-
2dpg	monomer (L)	2.5	0.232	0.173	Co-enzyme
1e7y	monomer (L)	2.48	0.296	0.205	G6P and structural $NADP^+$
1e7m	monomer (L)	2.54	0.299	0.222	-
1e77	monomer (L)	2.69	0.285	0.180	G6P
1h9b	monomer (L)	2.4	0.236	0.190	-
1h9a	monomer (L)	2.16	0.226	0.187	Co-enzyme
1h94	monomer (L)	2.5	0.292	0.213	Co-enzyme
1h93	monomer (L)	2.2	0.282	0.206	-
5aq1	dimer (T)	2.65	0.226	0.200	G6P and co-enzyme

TABLE 1.2: All the resolved structures of the G6PD enzyme. The resolution is the measure of the overall quality and reflects the level of detail of the structure (the lower the better). The R-factor represents the level of refinement of the model and describes the level of agreement between the crystallographic model and the original X-ray diffraction data. Values above 0.5 indicates poor quality models, while values around 0.2 are usually index of good quality. The R-free is the R-factor calculated on a small set of random coordinates that have not been included in the refinement process. This gives better and less-biased measure of the refinement progress. In the table, H=Human, L=*Leuconostoc mesenteroides* and T=*Trypanosoma cruzi*

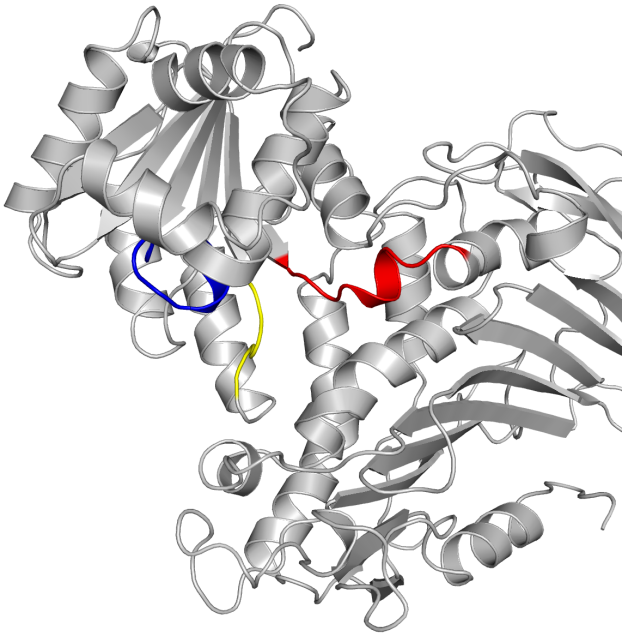


FIGURE 1.8: The G6PD dimer with the three highly conserved motives coloured in red (198-RIDHYLGKE-206), blue (38-GxxGDLA-44), and yellow (170-EKPxG-174).

G6PD is characterised by a two domain structure: an "NAD(P)-binding Rossmann-like" domain and a "Dihydrodipicolinate Reductase; domain 2" domain (Figure 1.9). The first domain is an $\alpha+\beta$ 3-layer(aba) sandwich structure (Figure 1.10a), commonly found in dehydrogenases and in G6PD has the role of co-enzyme binding. The second domain is an $\alpha+\beta$ 2-layer sandwich (Figure 1.10b), it is found in the core of G6PD and contains the dimerisation interface that interacts with the same region in the other unit. Although the residues in the latter region are not highly conserved, the geometry of the interface is maintained in all organisms. The residues found in the N-terminus are generally disordered and the formation of a disulphide bridge (C13-C446) prevents this area from moving too much, maintaining the catalytic activity [77]. An $NADP^+$ is located in a positively charged crevice between the β sheet and the C-terminus. Most of the $NADP^+$ structure lies inside the core of the enzyme, suggesting that its removal would affect both conformation and association of the subunits.

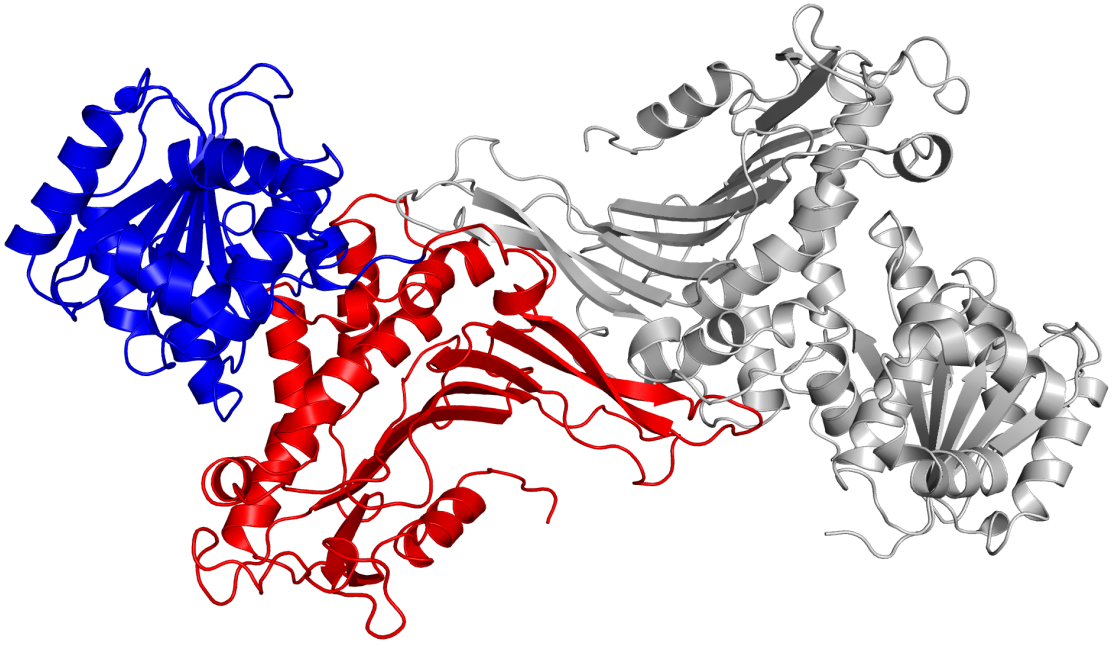
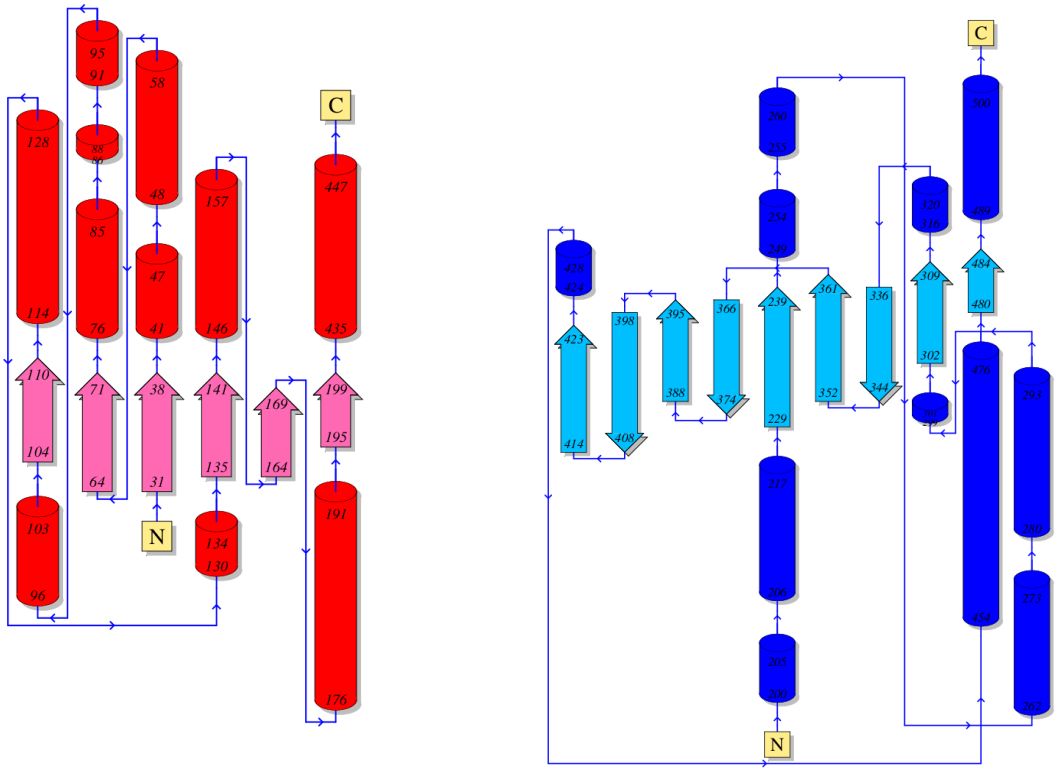


FIGURE 1.9: The G6PD dimer with the two domains coloured in only one chain. In blue, the "NADP-binding Rossmann-like Domain" (CATH code: 3.40.50.720) spans from V27 to I199 and from L433 to V453. In red, the central "Dihydrodipicolinate Reductase; domain 2" domain (CATH code: 3.30.360.10) spanning from D200 to K432 and R454 to G505. The other chain is in grey.



(a) (b)

FIGURE 1.10: Diagrams of the two domains in G6PD. (a) The "NADP-binding Rossmann-like Domain" (CATH code: 3.40.50.720) and (b) the "Dihydrodipicolinate Reductase; domain 2" domain (CATH code: 3.30.360.10). Both diagrams were obtained from PDBsum.

1.3.3.1 The glucose-6-phosphate binding site

The glucose-6-phosphate (G6P) binding site is well ordered and is located in the pocket between the two $NADP^+$ binding sites. Most of the residues that interact with G6P are conserved and correspond to those shown in Figure 1.11 [81]. Y202, H201 and K205 are part of the highly conserved G6P fingerprints 198-RIDHYLGKE-206. The catalytic function of the enzyme is fulfilled by H263, located in the centre of the site. Mutations K205R and K205T are known to affect G6PD k_{cat}, demonstrating the key role of this residue to the catalysis in humans [82]. The conserved lysine (K171) at the bottom of the site, precedes P172 in the sequence (EKPxG) and plays a fundamental role in guaranteeing the correct positioning of both the G6P and the co-enzyme. Although there is a direct interaction between the Gly395 side chain and the phosphate of G6P, this residue is not conserved in non human species.

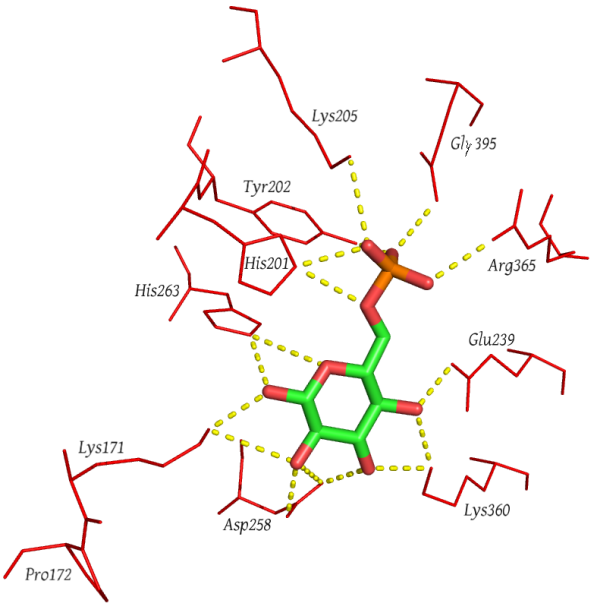


FIGURE 1.11: The hydrogen-bonding network in the substrate site.

1.3.3.2 The $NADP^+$ binding site

Similarly to the G6P binding site, the $NADP^+$ binding site is well ordered. The protein-$NADP^+$ interactions are presented in Figure 1.12. The adenine interaction environment is mediated by a triad of residues (G110-Q111-Y112), with the hydrogen bond with the oxygen of Tyr112 forcing the adenine ring to face the solvent. Moreover Leu142, Val146 and Pro143, while not directly involved in the binding, contribute to creating a hydrophobic surface for the adenine to bind.

1.3.3.3 The structural $NADP^+$ site

The structural $NADP^+$ site is found and conserved only in higher organisms and although no evidence has been collected, it may be because a long C-terminus allows a stronger bond with $NADP^+$. Prokaryotes have a shorter C-terminus than eukaryotes and therefore $NADP^+$ cannot bind. This site is important for the enzymatic activity of G6PD and is located between the β-sheet and C-terminus of the monomer (Figure 1.5 in orange). The $NADP^+$ is partially buried in a cleft (positively charged) on the protein surface (Figure 1.6 in orange) [73], without being strongly bound to the protein surface. Instead, the presence of a hinge region around G505 allows a dynamic equilibrium in which $NADP^+$ may migrate to the co-enzyme site in the presence of a low concentration of $NADP^+$. The C-terminus is a flexible structure, and when the tail moves, a group of negatively charged residues (E416, E417 and E419) is exposed to the surface of the

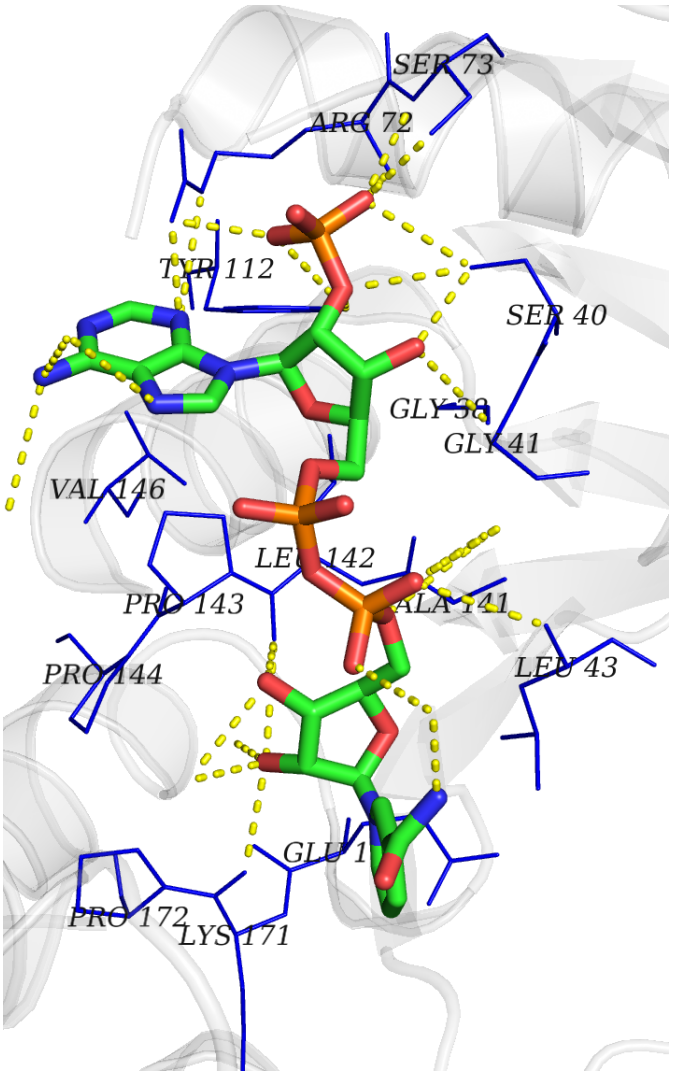


FIGURE 1.12: The hydrogen-bonding network in the $NADP^+$ binding site. The residues interacting with the $NADP^+$ are represented in blue.

$NADP^+$ binding site. This patch of residues only links the positively charged nicotinamide ring of $NADP^+$, prevents the linkage of NADPH when all the $NADP^+$ is reduced to NADPH. Another factor that may influence the binding at the $NADP^+$ structural site is substrate binding. Thanks to groups of residues that are close in the sequence, but span from one binding site to another, structure modification that occurs in one site can propagate and influence the behaviour of the other. These residues, implicated in the cross-connection, are:

1. **Arg365 and Lys366:** Arg365 binds the phosphate of G6P while Lys366 interacts with the phosphate of the structural $NADP^+$.

2. **Lys238 and Glu239:** Lys238 interacts with the structural $NADP^+$ and at the opposite edge of the strand, Glu239 forms hydrogen bonds with G6P.

The $NADP^+$ binding is extremely important and, not surprisingly, some of the class I G6PD variants (e.g. "Durham K238R", "Aachen" and "Loma Linda N363K") are clustered around this site.

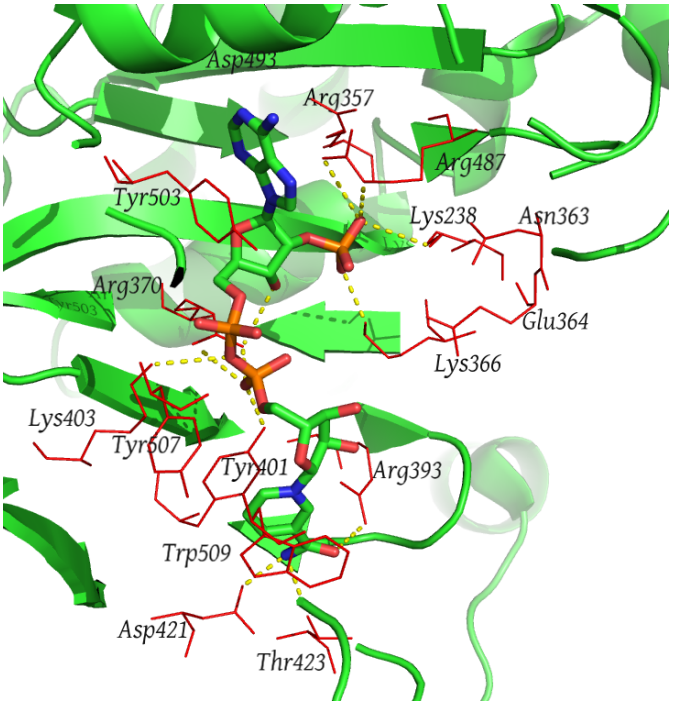


FIGURE 1.13: The hydrogen-bonding network in the structural $NADP^+$ binding site. The residues interacting with the $NADP^+$ are represented in red.

1.3.3.4 Proline 172

The conserved Pro172 (central residue of the peptide EKPxG) mediates the movement of the helix, allowing Lys171 to interact with both G6P through its terminal amino group and $NADP^+$ through the carbonyl group. Mutations in this amino acid exhibit class I deficiency, suggesting the importance of this residue in the correct positioning of both the substrate and the co-enzyme (Figure 1.14). In the tetramer structure, helix αb is longer than in the dimeric structure and the proline at position 172 (Pro172) is in *trans* in the tetramer while it is in *cis* in the dimer [73]. The *cis-trans* isomerisation of Pro172 could be an important mechanism for correct functioning of G6PD, but it could also be the result of crystallography artefacts or errors (the PDB files 2bhl has a resolution of 2.9 Å, while 1qki is 3 Å). During the experiments described in Chapter 3, all these hypotheses were considered, the state of Pro172 was monitored and a possible explanation is proposed.

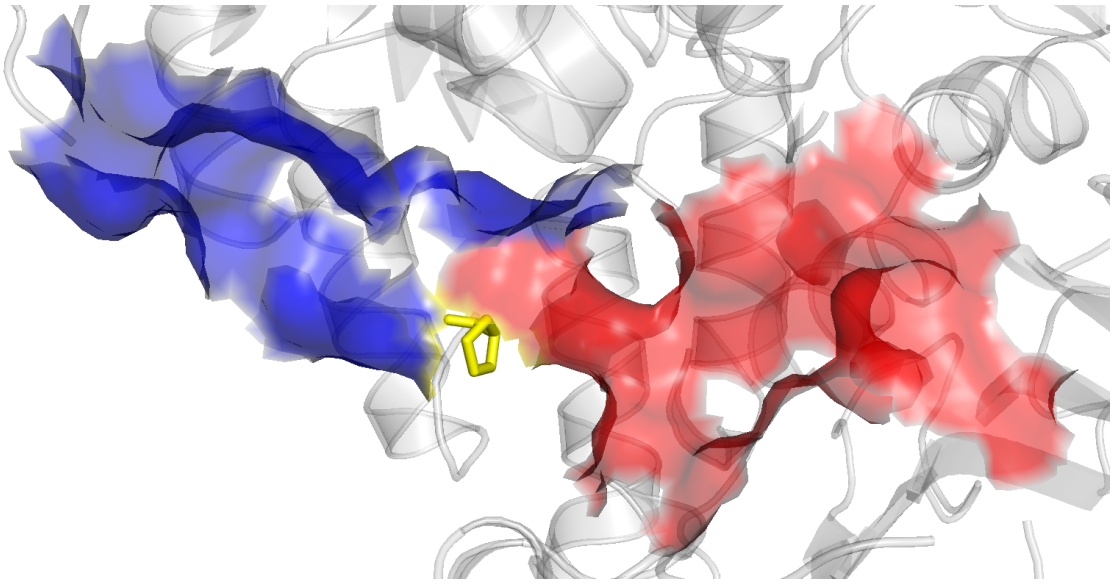


FIGURE 1.14: The central position of Pro172 (yellow), between the co-enzyme (blue) and the G6P (red) binding sites, mediates the correct positioning of both the substrate and the co-enzyme.

1.4 Thesis Overview

The overall aim of this thesis was to study the structure of G6PD and G6PD variants to provide information that could be used in the development of low-cost and simple-to-use immunological field tests for G6PD. The need for such tests comes from the strong relationship that exists between G6PD deficiency and the drugs used in malaria treatment. For individuals with G6PD depressed activity, ingesting certain malaria drugs (particularly the 8-aminoquinolines) may be dangerous as the G6PD-reduced activity greatly increases the side effects of these drugs. The starting idea of the project was that the depressed activity observed in the G6PD variants is the result of specific structural changes of the G6PD structure, that are markers to reduced, or normal activity. If this is true, it should be possible to develop a set of antibodies capable of binding specific features of the G6PD structure, differentiating between the wild-type and the mutants. These antibodies could then be used in the development of an immunological assay for G6PD variants similar to a pregnancy test kit. In summary, this thesis will describe how the joint use of the SAAP prediction methods and MD simulations, were used to study the behaviour of G6PD mutants in an attempt to find changes in the stability of the enzyme structure in the mutants. These changes should be detectable and similar among the mutants, but not the wild-type. To address this problem, data on the mutations and their effects were acquired using SAAPdap and SAAPpred. These tools are

capable of describing the effects of a given mutation and outline the mutations that are likely to affect the phenotype of a given enzyme.

- Chapter 2 introduces the computational methods used.
- Chapter 3 will present the results of extensive molecular dynamics simulations on both the wild-type and some of the G6PD variants. In this chapter, the mutants' behaviours are compared to the wild-type. Particular attention is given to the mechanisms that are capable of explaining the connection between structural features and phenotypic changes.
- Chapter 4 and 6 look at other approaches (metadynamics and network analysis) used better to understand how the effects of single point mutations are affecting the global structure of G6PD.
- Chapter 5 will present similar work to that presented in Chapter 2, but will do so using a coarse-grained force field: UNRES. The idea is further to increase the sampling to witness the big conformational changes that might be caused by the mutations.
- Chpater 7 provides a final discussion as well as conclusions and suggestions for future work.

Chapter 2

Computational background

Protein simulation *in silico* is an active field in modern biology, not only because it can help the understanding of the basic mechanisms of cell function, but also because it can radically change our approach toward diseases. In the last decades, the improvements in computer technology have allowed the birth of different approaches capable of handling the massive quantity of data generated by laboratories all over the world. With computer technology growing more powerful and affordable every year (e.g. GPUs), it becomes possible to model small molecules through detailed and precise methodology. Unfortunately, it is still not possible to use the same approach on larger system such as proteins or protein complexes. This does not mean that it is not possible to study those big systems, but a compromise between accuracy and practicality must be adopted. Over the years, different methodologies have been developed, all of them with the same idea in mind: reducing the complexity without losing details. With solid and well-developed roots, Molecular Mechanics (MM) is considered to be one of the most powerful.

2.1 Molecular Mechanics

Molecular modelling encompasses all the theoretical methods and techniques used to simulate the behaviour of biological molecules. For all these techniques, the minimum information required to describe a molecule is the location of the atoms of which it is composed. Unfortunately this assumption is in contrast with the quantum model of nature, but its formulation allows the study of large systems in an acceptable period of time. MM is a technique that makes use of the laws of classical mechanics to describe

the structure of molecules at equilibrium. The most informative description that can be given to a physical system is the wave function represented by the Schrödinger equation (Equation 2.1).

$$\mathrm{E}\Psi = \hat{\mathrm{H}}\Psi \tag{2.1}$$

Which indicates that at a stationary state (Ψ), the Hamiltonian operator ($\hat{\mathrm{H}}$) is the total energy (E) of that state. To determine the correct value of $\mathrm{E}\Psi$, all the atoms constituting the system (electrons, protons and neutrons) must be included in the calculations. To reduce the complexity of the system, it is generally more convenient to split equation 2.1 only into its nuclear and electronic components.

$$\Psi_{tot} = \Psi_{nuclei} + \Psi_{electrons} \tag{2.2}$$

Without going into details, the Born–Oppenheimer approximation (Equation 2.2) considers the electronic motion dependant only upon the nuclear position and not upon their velocities. Since the nucleus is much heavier in mass compared with the electrons, only motions of the nuclei are studied and the electrons are not explicitly examined [83]. The main advantage of this approach is that an atom can be simplified to a ball and the bonds in the molecules as springs connecting the balls together. This representation of molecules is at the base of modern force fields, and allow a lot of computational time to be saved.

2.1.1 Force Field

A force field is a mathematical function that estimates the energy value of a system, depending on its structural conformation. Force fields describe the forces acting on each atom of the system, allowing the calculation of the potential energy of a molecule as the sum of the single terms that a force field considers (i.e. bonded and non-bonded interactions). The basic formulation of a force field is the sum of the bonded terms, describing covalent bonds and the long-range non covalent forces. Formally:

$$E_{tot} = E_{bonded} + E_{non-bonded} \tag{2.3}$$

where the first term of the equation describes the deviation from an ideal geometry. Each term can be rewritten as a collection of simpler functions (2.29)

$$E_{bonded} = E_{bond} + E_{angles} + E_{dihedral} \tag{2.4}$$

$$E_{non-bonded} = E_{electrostatic} + E_{van-der-Waals} \tag{2.5}$$

2.1.1.1 Bonding stretching (E_{bond})

Many physical systems experience a linear restoring force when displaced from their equilibrium position. Thanks to the Born–Oppenheimer approximation (Equation 2.2) a diatomic molecule can be expressed by two spheres connected with a spring and thus described through Hooke's law, where k is the force constant for the spring and x is the deviation from ideal length.

$$F = -kx \tag{2.6}$$

The potential energy can be derived from equation 2.6, by assuming that it is equal to the negative of the force integrated over x, formally:

$$U(x) = -\int F dx = k\int x = \frac{1}{2}k(x)^2 \tag{2.7}$$

When internuclear distances are considered, equation 2.7 can be rewritten as follows:

$$U(x) = \frac{1}{2}k(r - r_0)^2 \tag{2.8}$$

where r and r_0 are the internuclear distances respectively at a certain distance (r) and at equilibrium (r_0). Although correct, these equations only partially represent reality because they do not consider bond breakage when the atoms are far away from each other. A more realistic model is represented by the *Morse potential* (Equation 2.9) that considers an energy cut-off beyond which rupture occurs.

$$U(x) = D_e\left(1 - e^{\beta(x-x_e)}\right) \tag{2.9}$$

In Equation 2.9, x is the distance between the atoms and x_e is the equilibrium distance, while D_e and β define the well depth and width of the potential. However the use of the harmonic oscillator instead of the Morse potential does not significantly reduce the accuracy of an experiment because at low energy conformations, the Morse potential and the harmonic oscillator have a very similar trend (Figure 2.1). Furthermore the

implementation of the Morse potential calculus is extremely computationally expensive and hence there is a preference for the use of the harmonic oscillator.

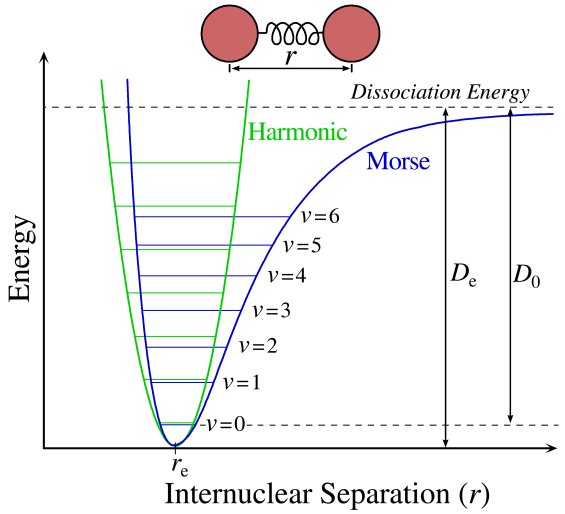


FIGURE 2.1: Ground-state potential energy curve for a diatomic molecule using the harmonic oscillator model (red) and the Morse potential (green). Figure by Mark Somoza, available under a Creative Commons Attribution-Noncommercial license.

2.1.1.2 Bending forces (E_{angles})

Similarly to what is done for bonding stretching, the energy variations associated with bond angle deformations are described by Equation 2.10:

$$U(\theta) = \frac{1}{2}k(\theta - \theta_e)^2 \tag{2.10}$$

where θ is the angle between three atoms and θ_e is the value at equilibrium.

2.1.1.3 Torsion forces ($E_{dihedral}$)

The torsion angle function models the motion associated with the rotation of dihedral angles, the angles between the two half-planes formed by four atoms, around the middle bond.

This potential is a periodic function with several minima, so can be modelled as a Taylor expansion of the cosine function (2.11), where ω and n are defined as the dihedral angle and the periodicy ($\frac{2\pi}{1...n}$).

$$U(\omega) = \frac{1}{2}k_\omega\Big[1 + cos(n(\omega - \omega_e))\Big] \tag{2.11}$$

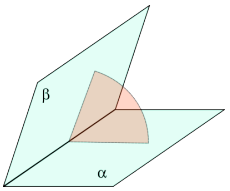


FIGURE 2.2: Dihedral angle between the half-planes α and β. Figure available under a Creative Commons Attribution-Noncommercial license.

These terms are generally different from one force field to another. The CHARMM force field, for example, has two additional terms; one is an interaction based on the distance between atoms separated by two bonds and the other is an improper dihedral term used to maintain chirality and planarity [84].

2.1.1.4 Electrostatic interactions ($E_{electrostatic}$)

The electrostatic interactions between a pair of atoms are the most complex forces to model because the potential fades very slowly with increasing distance. There are several implementations, but all of them are modelled on the Coulomb potential (Equation 2.12); $4\pi\epsilon_o$ is the dielectric function for the medium and r_{ij} is the distances between two charged atoms q_i and q_j.

$$E_{coul} = \sum_{i,j} \frac{q_i q_j}{4\pi\epsilon_0 \left|r_{ij}\right|^2} \tag{2.12}$$

2.1.1.5 Van der Waals forces ($E_{van-der-Waals}$)

The van der Waals interaction represents the balance between repulsive and attractive forces. The attractive force (power of 6), also called London's dispersion force, arises from the charge fluctuation in the electron cloud, while the repulsive force (power of 12) describes the repulsion of the electron clouds at close distance. When two atoms get closer, their electron densities begin to merge and, in the absence of bond formation, Pauli repulsion causes the energy to rise rapidly. These forces are well modelled (Figure 2.3) by the Lennard-Jones 6-12 potential (2.13).

$$E_{vdW} = \left[\left(\frac{A}{r^{12}}\right) - \left(\frac{B}{r^6}\right)\right] \tag{2.13}$$

with $A = 4\epsilon\sigma^{12}$ and $B = 4\epsilon\sigma^6$. These parameters are dependant on the atoms being considered. ϵ is the depth of the potential pitch, σ is the distance at which the potential

is zero and r is the distance between atoms.

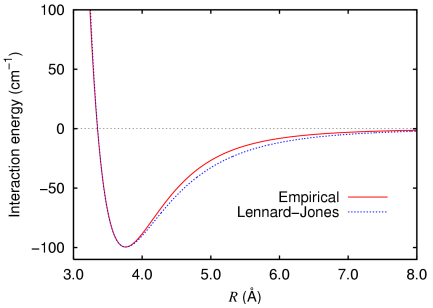


FIGURE 2.3: Van der Waals interaction energy between two argon atoms, as predicted by the Lennard-Jones potential (blue) and calculated by empirical measurements (red). Figure available under a Creative Commons Attribution-Noncommercial license.

The electrostatic interactions between hydrogen and very electronegative atoms, such as carbon or nitrogen, are not well described by standard atomic charges, so in the majority of force fields, hydrogen bonds are modelled implicitly as a combination of Couloumbic and Van der Waals terms. In the case of AMBER (Equation 2.15) this is achieved by ignoring the 12-6 interaction (hydrogen-acceptor) and adding a weak 12-10 potential.

2.1.1.6 Parametrization

The main goal of a force field designer is the development of a model that is as close as possible to experimental measurements. Because of the huge number of existing elements and their possible combinations, it is unlikely that there will be access to such a massive quantity of data. A common workaround is the use of arbitrary parameters to fit together experimental measurements and force field predictions. The deviation is calculated using a penalty function and minimized to match the experimental values better. Equation 2.14 is an example of such a function.

$$Z = \left[\sum_{i}^{Observables} \sum_{j}^{Occurrences} \frac{(calc_{i,j} - expti,j)^2}{w_i^2} \right]^{\frac{1}{2}} \tag{2.14}$$

Arbitrary parameters include explicit or implicit treatment of the atoms (all-atom *vs* united atom), the inclusion of solvation terms, specific water model optimization (for example TIP3P for both AMBER [85] and CHARMM force fields and SPC for the GROMOS force field [86]), long-distance cut-off and many more. Different force field use different parameters because they are designed for different purposes, using different measurements, approaches and optimisations. It is always good practice to avoid mixing together energy elements from different force fields.

The other main divergence between force fields is the number of order terms used to model the energy terms. Force fields that only use the energy terms described above are called *class I* additive potentials, as opposed to *class II* force fields that include higher order terms. These addictions are used to model a wider range of phenomena, such as bond breakage and cross terms (coupling of different internal variables). Although class II force fields are more accurate, a good parametrization allows the adequate treatment of most common large biomolecular systems using class I [87]. AMBER and CHARMM are two of the most-used force fields and their form is expressed in equations 2.15 and 2.16. Compared with AMBER, CHARMM has two additional terms; the Urey-Bradley (U_{UB}) and the improper dihedral terms. The U_{UB} harmonic term is an addition to the standard bond stretching and angle bending terms and models the interaction between atoms separated by two bonds (1,3 interaction), while the second helps in maintaining chirality and planarity.

AMBER *(Assisted Model Building and Energy Refinement)*

$$\begin{aligned} E_{tot} = &\sum_{bond} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedral} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \\ &+ \sum_{i<k}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\epsilon R_{ij}}\right] + \sum_{H_{bonds}}\left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right] \end{aligned} \tag{2.15}$$

CHARMM *(Chemistry at Harvard Macromolecular Mechanics)*

$$\begin{aligned} E_{tot} = &\sum_{bond} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedral} K_\phi[1 + \cos(n\phi - \delta)] + \\ &+ \sum_{impropers} K_\omega(\omega - \omega_{eq})^2 + \sum_{Urey-Bradley} K_u(u - u_{eq})^2 + \\ &+ \sum_{nonbonded} \epsilon\left[\left(\frac{R_{min_{i,j}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{i,j}}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\epsilon r_{i,j}} \end{aligned} \tag{2.16}$$

2.1.2 Potential Energy Surface (PES)

The best structure, from a chemical point of view, is defined as the structure with the lowest possible energy given a certain molecule. In physiological conditions, molecules do not exist in an isolated state, but rather in an equilibrium of a large number of different conformations; for this reason instead of considering a single conformation for a given chemical element, it is always better to think in terms of an ensemble of conformations for that specific molecule. The Potential Energy Surface (PES), is the hypersurface defined by the potential energy of a collection of atoms in all possible arrangements [88].

The position of all the atoms is defined by their coordinates (x,y,z), therefore, every point on the PES is described by the vector X:

$$X \equiv (x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_N, y_N, z_N) \equiv (\vec{r_1}, \vec{r_2}, \ldots, \vec{r_N}) \tag{2.17}$$

For a molecule with N atoms (N≥3) the PES has 3N-6 coordinate dimensions. Six dimensions are removed by imposing that the molecular center of the mass is at the origin and that the molecule is aligned with the axes. Along the PES, two types of states can be defined: *local minima*, which correspond to optimal low energy structures and *saddle points*, which are energy barriers on paths connecting minima. Unfortunately, the complete exploration of the PES is impossible, because of the great complexity of biological molecules. Typically only a small slice of surface is studied every time, but because only low energy states are explored, that slice can be considered as being informative (see section 5.4).

The main goal of MM is the search for the point of lowest energy, called the *global minimum*. *Optimization theory* is the field of applied mathematics that studies the problem of minimization of a function of several dimensions. Although there is no universal algorithm for optimization, all of them can be clustered into two families: *local search* and *global search*.

2.1.2.1 Methods of local search

These methods are analytic approaches that are able to find, in a reasonable time and with good approximation, the local minimum of a function. Because all of them use the gradient to move along the PES, they cannot climb energy barriers and they therefore stop once a minimum is reached (*downhill* approach).

1. **Steepest Descent**: For a random point of the function (x_n), this technique first calculates both the value of the function ($F(x_n)$) and the gradient (∇) at that point and then moves in the direction indicated by the negative of the gradient. The main idea is to identify the fastest way down to a local minimum. Formally:

$$x_{n+1} = x_n - a_n \nabla F(x_n) \tag{2.18}$$

with

$$F(x_0) \geq F(x_1) \geq F(x_2) \geq \ldots \tag{2.19}$$

Eventually the sequence (x_n) converges to the local minimum (Figure 2.4). The value of the step size (a_n) can be calculated in different ways; although computationally expensive, the formally correct way is simply to perform a "linear search" (proceeding by adjacent points) in the direction of the gradient.

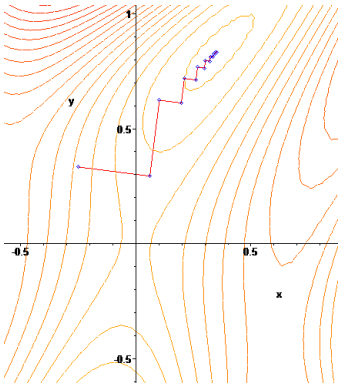


FIGURE 2.4: The steepest descent algorithm allow a function to converge in a point of minimum energy by changing direction following the negative of the gradient of the function in a point.

2. **Conjugate Gradient**: Conjugate gradient methods are very similar to steepest descent, with the only difference being that they orient the search direction by using a set of conjugate directions instead of only considering the gradient of the function at a given point. Generally, compared with steepest descent methods, conjugate gradient methods are computationally cheaper, but they tend to converge more slowly.

3. **Newton-Raphson**: This method is a root-finding algorithm that is able to locate a point of minimum energy by using the first few terms of the Taylor series [89]. Given a function $f(x)$, the Taylor series for the point $x = x_0 + \epsilon$ is:

$$f(x_0 + \epsilon) = f(x_0) + f'(x_0)\epsilon + \frac{1}{2}f''(x_0)\epsilon + \dots \tag{2.20}$$

The first order terms (Equation 2.21) represent the equation of the tangent line to the curve at $(x_0, f(x_0))$ and $(x_1, 0)$ is the place where the tangent line intersects the x-axis.

$$f(x_0 + \epsilon) \approx f(x_0) + f'(x_0)\epsilon \tag{2.21}$$

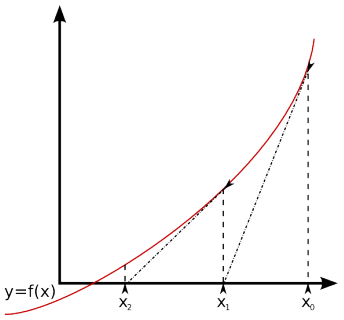


FIGURE 2.5: The Newton-Raphson algorithm works by calculating the intersection to the x-axis of the tangent line of a point, and proceed until the value of the tangent line is equal to 0.

By setting $f(x_0 + \epsilon) = 0$, the step needed to land closer to the minimum from an initial position x_0 can be inferred from Equation 2.22.

$$\epsilon_0 = -\frac{f(x_0)}{f'(x_0)} \tag{2.22}$$

At every step, a new ϵ is calculated until convergence to a local minimum is reached. The final implementation that is used iteratively in the Newton-Raphson method is the following:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{2.23}$$

2.1.2.2 Methods of global search

Despite their name, these methods do not guarantee that the global minimum will be found, but they allow the exploration of a bigger slice of the PES compared with a local search.

1. **Simulating Annealing**: Stochastic models based on the Monte Carlo sampling (probability accumulation), derive from the annealing proces in metallurgy, where a metal is heated and slowly cooled down to improve the quality of an alloy. Given a state (point on the PES) at a certain temperature (T), a *Markov chain* is used to represent all the states that have a transition probability (p) of being chosen as the next state. If $\Delta E < 0$ the transition occurs because a state of lower level energy has been found; however if $\Delta E > 0$ the transition may occur with p probability.

$$p = exp\left(-\frac{\Delta E}{k_b T}\right) \tag{2.24}$$

The transition probability is a function of the Boltzman constant (k_b) and the temperature of the system. The latter plays a crucial role because it controls the evolution of the system itself. At low temperature (low energy of the system), there is a low probability of exit from a potential well ($\Delta E > 0$) and, as a consequence, the Markov chain will converge to a minimum. The general approach to simulating annealing is to fix an initial temperature together with a cooling schedule that will slowly reduce the energy until a huge slice of PES has been explored and a minimum is found. It is important to understand that the minimum found by this method is not guaranteed to be the global minimum, but it is a good approximation of it.

2. The **Genetic Algorithm** is a heuristic method that draws its origin from the field of population genetics and tries to simulate the process of natural evolution. The main idea is that a population of different solutions to a problem (a function), tends naturally to evolve to the best solution(s). Genetic algorithms make use of two forces, natural selection and sexual reproduction, to push the evolution process towards convergence. Different implementations of this algorithm exist, but all of them share the same rationale. That is:

   (a) *start*: generate a random population with n chromosomes;

   (b) *fitness*: calculate the fitness: $f(x)$ (e.g. potential energy);

   (c) *new population*: generate a new population in three steps:

      i. based on the fitness, a pair of parents are chosen;

      ii. a random pair of parental chromosomes is generated;

      iii. random mutation events occur with a given probability;

   (d) *test*: verify the stop conditions;

   (e) *iteration*: if the stop conditions are not reached, the algorithm returns to step (b).

2.2 Statistical thermodynamics

Classical thermodynamics studies nature with a macroscopic approach, where the system is considered at thermodynamic equilibrium and the atomic detail is completely neglected. All the equations that describe the properties of a system can be written using only four variables: Temperature (T), Pressure (P), Volume (V) and Number of moles (N). Thermodynamics reproduces the system as an *ideal gas*, with the result that objects are described as point particles with elastic collisions. The statistical thermodynamics approach, on the other hand, tries to increase the level of information by considering the atomistic nature of the system. As a direct consequence, all the properties must be rewritten to account for the behaviour of an ensemble of molecules in the order of the Avogadro constant, N_A. In statistical thermodynamics, a generic macroscopic property, A, is described as a function of the coordinates ($\vec{r}$) and the momenta ($\vec{p}$) of all the particles of the system.

2.2.1 Boltzmann distribution

In a system consisting of N_{tot} particles (with N_{tot} in the order of Avogadro number), all the particles belong to a specific state with energy E_U. At $T = 0K$ all the particles of the system are at the ground state (global minimum) with $E_U(r_0) = 0$, where r are the coordinates of the system. Statistical thermodynamics tries to understand the probability that a certain energy state is populated. If a general state i, with $E_{U,i} > E_{U,0}$, is populated according to Equation 2.25

$$\frac{N_j}{N_{tot}} > \frac{N_i}{N_{tot}} \Rightarrow N_j > N_i \tag{2.25}$$

where N_i is the number of particles at the E_i state, and

$$1 > \frac{N_i}{N_{tot}} > 0 \tag{2.26}$$

then the state N_j is more populated than N_i, and it is therefore more probable to find a particle there. The *Boltzmann distribution* (Equation 2.27) defines the temperature (T) dependent function which assigns the probability of being populated to every state of the system.

$$\frac{N_i}{N_{tot}} = \frac{e^{-\frac{E_i(ri)}{k_b T}}}{\sum_i e^{-\frac{E_i(ri)}{k_b T}}} = \frac{e^{-\frac{E_i(ri)}{k_b T}}}{Z(T,r)} = \sigma(E_i(r), T) \tag{2.27}$$

k_b is the *Boltzmann constant* and represents the proportionality factor between entropy (S) and the number of the accessible states (Ω) (Equation 2.28).

$$S_i = k_b \ln(\Omega) \tag{2.28}$$

$Z(T, r)$, also called normalizing function, is the repartition function that contains all the information which is necessary to understand the system, since it gives all the energy values of all the possible states. σ is a probability distribution and it can only assume values between 0 and 1. The Boltzmann distribution exhibits the following properties:

1. The ground state is always the most populated;
2. A high energy state is always less populated than a low energy state.
3. The distribution is in accordance with the third principle of thermodynamics: The entropy of a system depends on the temperature (T). If $T = 0$, the entropy is 0 and only one system is populated, but a $T \to \infty$ then all the states will become equally populated.

2.2.2 Phases space and Ergodic hypothesis

The total energy of a system is the sum of both kinetic and potential contributions (Equation 2.29). The previous sections explained how a force field connects potential energy and molecular conformation together (Equation 2.1.1).

$$E(p, r) = E_{kinetic}(p) + E_{potential}(r) \tag{2.29}$$

$$E_{kinetic}(p_i) = \frac{p_i^2}{2m_i} \tag{2.30}$$

Because kinetic energy is a function of mass (m) and momentum (p) (Equation 2.30), the total energy depends only on the positions and on the momenta of the particles of the system. This leads to the definition of the *phase space*: for a system of N particles, the phase space is the mathematical space ($\mathbf{X}$) in which all possible conformations (r) and momenta are represented.

$$\mathbf{X} = \{p, r\} | \mathbb{R}^{6N} = \mathbb{R}^{3N}(r) \times \mathbb{R}^{3N}(p) \tag{2.31}$$

with

$$\begin{aligned}\vec{p} =&\{p_{x1}, p_{y1}, p_{z1}, p_{x2}, p_{y2}, p_{z2}, \ldots, p_{xN}, p_{yN}, p_{zN}\}\\ \vec{r} =&\{x_1, y_1, z_1, x_2, y_2, z_2, \ldots, x_N, y_N, z_N\}\end{aligned} \tag{2.32}$$

where x,y,z represent the coordinates. Inside the phase space and using the Boltzmann distribution, it is possible to rewrite any macroscopic properties and a generic function $A = f(\vec{r}, \vec{p})$ becomes

$$A = \iint A(p, r)\sigma(r, T)\, dp\, dr \tag{2.33}$$

If the property studied is the energy of the system, the result will be expressed by

$$\langle A\rangle_{ensemble} = \iint A(p, r)\sigma(E(r), T)\, dp\, dr \tag{2.34}$$

Energy, speed, pressure and volume are all continuous variables (here the integral) that require an infinite time to be fully described. However, one can experimentally sample a certain property over time (t), and obtain an estimation of the average values:

$$\langle A\rangle_{ave} \approx \frac{1}{M_{step}} \sum_{i=1}^{M_{step}} A(p(t_i), r(t_i)) \tag{2.35}$$

with M equal to the number of steps of the experiment. Although equation 2.35 only partially describes A, if the number of steps approaches infinity ($M \to \infty$), $\langle A\rangle_{time}$ becomes a good approximation of $\langle A\rangle_{ensemble}$:

$$\langle A\rangle_{time} = \lim_{M\to\infty} \frac{1}{M} \int_{t=0}^{M} A(p(t), r(t))dt \tag{2.36}$$

The fundamental principle in which $\langle A\rangle_{ensemble} = \langle A\rangle_{time}$ is called the *ergodic hypothesis*. This hypothesis implies that the system goes through every spatial and kinetic configuration during its temporal evolution. It is not possible to calculate this continuous function, but numerical integration techniques such as Molecular Dynamics (MD) and Monte Carlo (MC), are able to approximate the solutions calculating a trajectory, r_i (Equation 2.37), in the phase space.

$$\left\{\begin{array}{l}\vec{r_i} = \vec{r_i}(t),\ \forall i\\ \vec{p_i} = \vec{p_i}(t),\ \forall i\end{array}\right. = \left\{\begin{array}{l}\vec{r_i} = \vec{r_i}(t_o), \vec{r_i}(t_1), \ldots\\ \vec{p_i} = \vec{p_i}(t_o), \vec{p_i}(t_1), \ldots\end{array}\right. \tag{2.37}$$

These techniques sample, for a sufficiently long time, the phase space in the informative low energy regions of the Boltzmann distribution.

2.3 Molecular Dynamics

Molecular dynamics (MD) is a system of numerical integration that tries to understand the motion of the atoms of molecules in a well-defined time frame. Using physiological conditions in the simulation, an MD experiment is able to study, *in silico*, the low level energy structures that are responsible for the *in vivo* behaviours.

Every atom (i) is subject to a force according to the classical Newtonian definition:

$$\vec{F} = m_i \vec{a_i} \tag{2.38}$$

but if the forces are conservative (the value depends only on the position), (Equation 2.38) becomes

$$\vec{F} = m_i \vec{a_i} = m_i \frac{d^2 \vec{r_i}(t_i)}{dt^2} = -\frac{dE(\vec{r_i})}{d\vec{r_i}}, \forall i \tag{2.39}$$

or

$$\vec{F} = -\frac{dE(\vec{r_i})}{d\vec{r_i}} = -\nabla E(\vec{r_i}) \tag{2.40}$$

Equation 2.40 suggests that the force is the negative gradient of the potential energy ($-\nabla E(\vec{r_i})$), meaning that an atom tends to move in the direction opposite to the direction of the energy function growth. Therefore, at a point of minimum energy, the gradient is zero and the atom maintains its position. By integrating the forces which act on every atom of the system, it is possible to express the relationship between potential energy and position over time:

$$\begin{cases} \vec{r_i} = \vec{r_i}(t), \forall i \\ \vec{p_i} = \vec{p_i}(t), \forall i \end{cases} \tag{2.41}$$

This system of equation Equation 2.41 is solved using the Verlet (leapfrog) algorithm, or one of its variants.

2.3.1 Verlet integration

All the numerical integration algorithms use a truncated Taylor series of both the positions (r) and the momenta (p) of the atoms of the system.

$$\begin{cases} r_i(t+\delta t) = r_i(t) + r_i'(t)\delta t + \frac{1}{2}r_i''(t)\delta t^2 \\ p_i(t+\delta t) = p_i(t) + p_i'(t)\delta t + \frac{1}{2}p_i''(t)\delta t^2 \end{cases} \tag{2.42}$$

Given a starting configuration (t) the algorithm calculates the values of the conformation after a small increment of time ($t + \delta t$).

$$r_i(t+\delta t) = r_i(t) + r_i'(t)\delta t + \frac{1}{2}r_i''(t)\delta t^2 = r_i(t) + v_i(t)\delta t + \frac{1}{2}a_i(t)\delta t^2 \tag{2.43}$$

where $v_i(t)$ is the vector of the velocities at step t and $a_i(t)$ is the vector of the accelerations at the same step (t). Because t can be calculated as the sum of both $t + \delta t$ and $t - \delta t$, (Equation 2.43) can be rewritten as

$$\begin{aligned} r_i(t+\delta t) + r_i(t-\delta t) &= r_i(t) + v_i(t)\delta t + \frac{1}{2}a_i(t)\delta t^2 + r_i(t) - v_i(t)\delta t + \frac{1}{2}a_i(t)\delta t^2 \\ &= 2r_i(t) + a_i(t)\delta t^2 \end{aligned} \tag{2.44}$$

Using the gradient of the potential energy (Equation 2.39), the acceleration is replaced by the force, obtaining

$$2r_i(t) + \frac{F_i}{m_i}\delta t^2 \tag{2.45}$$

and finally

$$r_i(t+\delta t) = 2r_i(t) - r_i(t-\delta t) + \frac{F_i}{m_i}\delta t^2 \tag{2.46}$$

where $r_i(t+\delta t)$ is the term referring to the new step, $2r_i(t)$ is the term referring to the previous step, while $r_i(t-\delta t)$ refers to the second previous step.

The integration of the momenta is treated slightly differently. The first velocities (p_0) are randomly generated from the Boltzman distributions of velocities:

$$v_i(t=0) = \left(\frac{m_i}{2\pi k_b T}\right)^{\frac{1}{2}} e^{-\frac{m_i v_i^2}{2k_b T}} \tag{2.47}$$

whith k_b, m and v being the Boltzman constant, the mass and the velocities respectively.

In statistical thermodynamics the *equipartition theorem* states that, at equilibrium, the total kinetic energy (K) is shared equally among all of its parts. Equation 2.48 expresses this relationship for a system with one degree of freedom.

$$K = \frac{1}{2} k_b T \tag{2.48}$$

When the system is composed of N atoms, with 3 degrees of freedom each, Equation 2.48 becomes

$$K = \sum_{i=1}^{N} \left(\frac{3}{2} k_b T \right) = \frac{3}{2} N k_b T \tag{2.49}$$

During an MD simulation, the system is described using a classical mechanics point of view, which allows easy calculation of the momenta (p_i) imposing:

$$K = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} = \frac{3}{2} N k_b T \tag{2.50}$$

Once some initial coordinates (a PDB file), some random velocities and the temperature are defined, it is possible to describe fully the evolution of a dynamic system. However, to maintain a correct description of the dynamics, the system must be isolated (N is constant). To achieve this, three different *ensembles* can be used:

- NVE (*micro canonical ensemble*): volume and energy are maintained constant;
- NVT (*canonical ensemble*): temperature and volume are maintained constant;
- NPT (*isothermal–isobaric ensemble*): temperature and pressure are maintained constant;

The temperature is maintained constant using a multiplicative factor (λ) that is re-calculated at every step. Because the system evolves over time, the kinetic energy at $t + \delta t$ differs from the values recorded in the previous step (K_t) and is calculated from Equation 2.50

$$K_{t+\delta t} = \sum_{i=1}^{N} \frac{(\lambda p_i)^2}{2m_i} = \frac{3}{2} N k_b T_{t+\delta t} \tag{2.51}$$

If

$$\Delta K = K_{t+\delta t} - K_t = \sum_{i=1}^{N} \frac{(\lambda p_i)^2}{2m_i} - \sum_{i=1}^{N} \frac{p_i^2}{2m_i} = \frac{3}{2} N k_b (T_{t+\delta t} - T_t) \tag{2.52}$$

and if

$$p_i(t + \delta t) = p_i(t) * \lambda \tag{2.53}$$

then

$$\lambda = \sqrt{\frac{T_{t+\delta t}}{T_t}} \tag{2.54}$$

The volume is maintained constant using *periodic boundary conditions* (Figure 2.6), in which the system is contained in a simulation box surrounded by an infinite number of completely equal boxes. The boxes intercommunicate, and if a particle exits the simulation box from one side, it will re-enter from the opposite side.
For details on the algorithm used in this project, refer to Chapter 3 and Chapter 5.

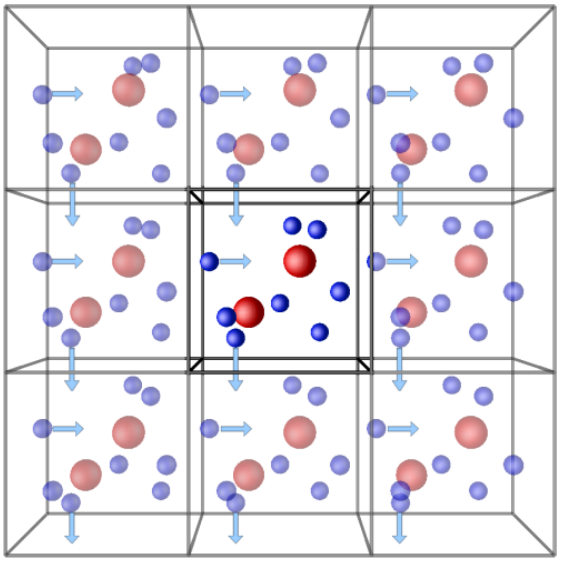


FIGURE 2.6: Representation of the periodic boundary conditions. The simulation box (centre box) is surrounded by identical boxes and every time a particle leaves the simulation box, the same particle re-enters from an adjacent box.
Figure from *isaacs.sourceforge.net/phys*

2.3.2 The Sampling problem

Molecular dynamics models are a cheap and reliable tool to explore the conformational space of biomolecules. Unfortunately the usefulness of MD methods is limited by the quantity of PES that can be explored during a simulation: the sampling. The PES is a complex hypersurface which describes a molecule and all its possible states and interconnecting pathways. Because of its dimensions and complexity it is impossible to describe its topology in fullness, and MD experiments are only capable of exploring small slices of it. The resulting uncertainty is problematic because MD yields incomplete information, which affects both the accuracy and the reproducibility of the experiments. A common way of tackling this problem is the use of several short simulations, that considered as a whole, can give a more correct description of the system studied. This approach is fine for most molecules, but there are cases in which longer simulations are required: protein folding and unfolding for example. In such cases, one of the ways used to improve the sampling is to raise the temperature of the simulation. Equation 2.48 describes how the kinetic energy (K) is linked to the temperature, and suggests that variations in temperature affect the internal energy of the molecules in the system. The additional energy makes changes of states more probable, and therefore, the number of conformations explored during an experiment is increased. The more the temperature rises, the larger the sampling will be. The drawback is that, at higher temperature, the probability of exploring biologically important structures decreases. Techniques such as 'replica exchange molecular dynamics' (REMD) overcome the problem by running several independent replicas at different temperatures and exchanging conformations among trajectories, allowing low energy structures to sample conformations that would not be explored otherwise. REMD simulations guarantee a massive increase in sampling, while exploring structures biologically relevant to the problem studied. Other ways of improving the sampling include the use of coarse grained models and metadynamics simulations. The former consists in increasing the simulation performance by averaging the energy over certain degrees of freedom, while the latter adds Gaussian functions to the potential of some selected properties to fill the potential well. Generating a correct sampling was a major problem during the G6PD simulations, and Chapter 5 and Chapter 6 will describe how the problem was treated.

2.4 SAAP

Predicting the effects of mutations on protein phenotype is an important field of research. Several serious diseases, such as Cystic fibrosis, neurofibromastosis, colour blindness and sickle-cell anaemia, are caused by deleterious phenotypic changes resulting from single amino acid changes having an effect on protein structure. The SAAP (Single Amino Acid Polymorphism) resources [90–93] are tools capable of explaining the possible effects of mutations by mapping mutations to the protein structure and analysing their likely effects. Single Nucleotide Polymorphisms (SNPs) are mutations that occur in at least 1% of a normal population. In the context of analysing the effects of mutations the term is often used for mutations which have no apparent effect on phenotype. In contrast, Pathogenic Deviations (PDs) are low frequency mutations that are capable of causing disease. Over time several methods aiming to analyse and predict the damaging effects of mutations have been developed. Some methods are sequence based and calculate conservation scores from multiple alignments (e.g.SIFT [94], PANTHER [95] and MutationAssessor [96]), others use machine learning on sets of both sequence and structural data from known structures of predicted properties, such as solvent accessibility (PolyPhen-2 [97]), and other combine the outputs from other predictors (Condel [98]). Because each method has a different definition of the 'boundaries' between SNPs and PDs, it is not easy to compare their performance directly, but their Matthews' correlation coefficient (MCC) was found to vary from 0.453 to 0.671 when tested on the HumVar dataset (all human disease-causing, other than cancer, mutations from UniProtKB) [98]. Even though the SAAP tools are limited by the necessity of having a resolved protein structure to work, SAAPpred, trained and tested on the same dataset and using 10-fold cross-validation, has a MCC of 0.692, outperforming the competing methods. The analysis of a mutation using the SAAP tools starts with the SAAPdap pipeline [93]. The idea is to understand the likely structural effects that the mutation has on the protein structure, comparing known effects in SNPs and PDs. This is achieved by mapping the SAAPs on the 3D structure (PDB structure required) and performing fourteen different structural analyses, such as checking for the introduction of destabilising voids in the structure, steric clashes with existing residues and the introduction of unfavourable torsion angles (refer to Table 2.1 for a complete description). The collected data are presented in both an overview mode (Figure 2.7) and a detailed one (Figure 2.8). Because the pathogenic phenotype could be the result of numerous, but small, changes to the protein structure, the SAAPdap results are further analysed by SAAPpred [93].

SAAPpred is a 47-feature machine learning method that uses Random Forests [99], implemented in Weka [100], to predict whether the mutation is a SNP or a PD. SAAPpred results use a confidence score system (from 0 to 1) that indicate how likely the prediction is to be correct. Values above a fixed threshold (> 0.1) indicate mutations that are likely associated with severe phenotypic alteration.

Analysis	Description
Interface	Residue is in an interfece according to solvent accesibility criteria;
Binding	Residues maked interactions with a different protein chain or ligand;
SprotFT	Residue is annotated as functional relevant by UniProtKB/SwissProt;
Clash	Mutation introduces a steric clash with an existing residue;
Void	Mutation introduces a destabilizing void in the protein core;
Cis-Proline	Mutation from cis-proline, introducing an unfavourable omega torsion angle;
Glycine	Mutation from glycine, introducing an unfavourable omega torsion angle;
Proline	Mutation from proline, introducing an unfavourable omega torsion angle;
HBond	Mutation disrupts a hydrogen bond;
Corephilic	Introduction of a hydrophilic residue in the protein core;
Surfacephobic	Introduction of a hydrophobic residue on the protein surface;
Buriedcharge	Mutation causes an unsatisfied charge in the protein core;
SSgeometry	Mutation disrupts a disulphide bond;
Impact	Residue is significantly conserved.

TABLE 2.1: List of the SAAPdap analyses as listed in *Al-Numail and Martin* [93].

N 226 -> S	11	No structural effects identified	[JSON]
N 226 -> D	11	No structural effects identified	[JSON]
N 226 -> I	11	HBonds (10) | SurfacePhobic (6)	[JSON]
N 226 -> K	11	HBonds (9)	[JSON]
R 227 -> W	11	Binding (5) | Interface (8) | SurfacePhobic (11)	[JSON]
R 227 -> G	11	Binding (5) | HBonds (11) | Interface (8)	[JSON]
R 227 -> L	11	Binding (5) | HBonds (11) | Interface (8) | SurfacePhobic (11)	[JSON]
R 227 -> P	11	Binding (5) | Clash (11) | HBonds (11) | Interface (8)	[JSON]
R 227 -> Q	11	Binding (5) | Interface (8)	[JSON]
D 228 -> V	11	Interface (8) | SurfacePhobic (11)	[JSON]
D 228 -> Y	11	Interface (8)	[JSON]

FIGURE 2.7: Example of SAAPdap output (overview mode). For each mutation (1st column) all the likely structural effects are listed (3rd column), together with the number of resolved structures (PDB files) used for the analyses (2nd column). The sign in parenthesis in column 3 are the numbers of structures in which the given effect is predicted to occur. The last column presents a link (JSON) to a file which stores all the information for that specific mutation. The colour gives an indicator, from green to red, of the likelihood of a damaging effect based on the fraction of structures affected.

UniProt Entry: P11413
Mutation: G 306->R

Summary

PDB	chain	HBonds	Binding	BuriedCharge	Voids	SProtFT	SurfacePhobic	Interface	Glycine	Clash	CisPro	Proline	CorePhilic	Impact	SSGeom
1qki	A			X						X			X	X	
1qki	B			X		?				X			X	X	
1qki	C			X						X			X	X	
1qki	D			X						X			X	X	
1qki	E			X						X			X	X	
1qki	F			X						X			X	X	
1qki	G			X						X			X	X	
1qki	H			X						X			X	X	
2bh9	A			X						X			X	X	
2bhl	A			X						X			X	X	
2bhl	B			X						X			X	X	

Hover over the column titles for an explanation.

Structural analysis

Expand all | Collapse all

1qki Residue A306 | Structure type: crystal | Resolution: 3.00A | R-factor: 24.70%

HBonds No problems identified	Binding No problems identified
BuriedCharge The mutation resulted in introducing or removing a buried charge.	Voids No problems identified
SProtFT No problems identified	SurfacePhobic No problems identified
Interface No problems identified	Glycine No problems identified
Clash The replacement sidechain leads to a clash with surrounding residues. **The clash energy was 1650.58 kcal/mol.** 99% of sidechains in real proteins have an energy less than 34.33 kcal/mol. Consequently energies >34.33 kcal/mol and <50 kcal/mol can be considered mild clashes, 50-100 kcal/mol medium clashes, >100 kcal/mol severe clashes. Note that clash energies can be extremely high (>>100000 kcal/mol)	CisPro No problems identified
Proline No problems identified	CorePhilic The mutation introduces a hydrophilic residue into the core of the protein.
Impact The mutation was at a site flagged as 'highly conserved'.	SSGeom No problems identified.

FIGURE 2.8: Example of SAAPdap output (detailed mode). Results are summarised at thte top where the effects on each known structure are presented. Below, all the structural analyses performed are presented and these can be expanded to provide mode details

Chapter 3

All-atom simulations

3.1 Overview

This chapter will describe how a set of G6PD mutations was selected and their effects were studied using MD all-atom simulations. The starting idea of the project was that the reduced catalytic activity observed in the mutants is caused by a large change in the equilibrium between correctly folded and misfolded states, where these states are significantly different from the correctly folded state. Initially, from the sequences of all the possible G6PD mutants, a subset of mutants was extracted. Because the total number of mutants was close to 3500, it was not possible to study all of them. Some rules were therefore set, to select only a representative sample of these mutations. The criteria used guaranteed that the mutations chosen were spread along the sequence, had different predicted structural effects, some of which were known pathogenic mutations. 17 mutants were identified and studied. In order to find a point in which there was a clear difference in behaviour between the wild-type and mutants, it was decided to start MD simulations at 310 and 500 K only. At 310 K it would have been possible to observe the protein behaviour at physiological conditions, while at 500 K the system would be expected to have enough energy to cause the complete unfolding of G6PD. In these conditions the enzyme stability was assessed and both temperatures, one being too low and the other too high, were found not to be ideal to explore the G6PD conformational space extensively. Additional temperatures were then explored in an attempt to find a temperature at which the mutants were clearly behaving differently from the wild-type. Therefore, simulations at 400 K, 450 K and 470 K were performed. All the data collected

indicated 400 K to be the best temperature to detect G6PD instability behaviours in the mutants.

3.2 Methodology: mutant selection

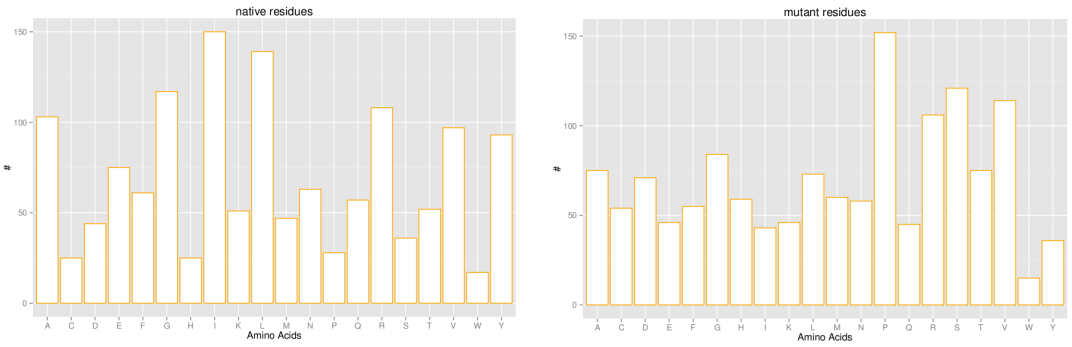
At the beginning of the project, around 186 existing G6PD variants were known [101], but to avoid missing important cases, all possible mutations of the G6PD gene resulting from single base changes were considered. Many mutations lead to mild to minor enzymatic deficiency, meaning that is only when certain malaria drugs or foods are ingested that the effects of the mutations are triggered. For this reason, there may be individuals who are unaware of their condition, because they have never used those drugs or eaten these foods. Initially the DNA sequence, obtained from the "*human metabolome database*" (*http://www.hmdb.ca/proteins/5564*) was mutated *in silico* to generate a list of possible mRNAs, through single point mutations of the original sequence. The resulting sequences were first translated and then filtered to eliminate duplications. Of the initial 4645 sequences, 3216 reached this point. Looking at the numbers, it appeared impossible to study all of them with MD simulations, so the next step was to reduce this number to a few dozen only. SAAPdap [93] was used to understand the possible structural effects resulting from each mutation event. For each mutant, the generated information was combined and processed by SAAPpred [93]. SAAPpred, through a machine learning algorithm, was able to predict how likely the mutation is to affect the phenotype of the protein. The extracts shown in Figure 2.7 and Figure 2.8 are examples of the types of information obtained from SAAPdap for G6PD. The complete results are accessible at the following address: *http://www.bioinf.org.uk/people/francesco/*.
The final step of the selection process was to pick, from the list of SAAPpred results, the most interesting and damaging mutations. This was done by setting the following criteria:

- **The mutations must have a high confidence of being damaging**;
- **The mutations must be spread throughout the protein structure**: This criterion was set to avoid studying only certain sections of the protein structure. If all the studied mutations were clustered around the same features (e.g. binding sites), it would be difficult to generalise the results, in the sense that the observed behaviours may be determined more by the position of the mutant than by the mutation itself;

- **Different residue types must be considered**: This criterion was introduced later on in the project. It was observed that the most damaging mutations concerned some residues more than others (arginine and proline in particular, see Figure 3.1). The introduction of this new criterion, allowed an increase in the range of cases and the study of less represented mutations;

- **Some mutations with known clinical effects must be included in the study**: The last criterion was critical because it helped validating the results obtained using real data. Observing that some mutants with known depressed activity present structural instability may validate the hypothesis that the depressed activity of G6PD mutants is owing to some structural damage in their structures.

Table 3.1 lists the mutants that fill the criteria and were therefore studied.



(a) (b)

FIGURE 3.1: (a) Native residue frequency and (b) mutant residue frequency of the mutants that were predicted damaging by SAAPpred. From the list of damaging mutants, The figures indicate that alanine, glycine, isoleucine and arginine are the most replaced residues in damaging G6PD variants, while proline, arginine, serine and valine are the most common replacement residues.

TABLE 3.1: List of the mutants studied with all-atom molecular dynamics simulations. All the mutants were predicted as damaging by SAAPpred (confidence is in brackets), with the exception of N128D. Some of the mutants are known G6PD variants (the type is in brackets). The red box indicates the two mutations constituting A^- : the only multi missense mutant studied.

Position	From	To	Predicted	Known variant
204	Glycine	Arginine	✓ (0.8)	-
306	Glycine	Arginine	✓ (0.8)	-
359	Glycine	Arginine	✓ (0.79)	-
264	Leucine	Arginine	✓ (0.79)	-
137	Leucine	Proline	✓ (0.78)	✓ (II)
140	Leucine	Proline	✓ (0.78)	-
338	Alanine	Glutamate	✓ (0.75)	-
370	Arginine	Tryptophan	✓ (0.71)	-
287	Glutamate	Lysine	✓ (0.69)	-
70	Cysteine	Tyrosine	✓ (0.67)	✓ (II)
306	Glycine	Serine	✓ (0.64)	✓ (II)
232	Cysteine	Tyrosine	✓ (0.49)	✓ (I)
269	Cysteine	Tyrosine	✓ (0.49)	✓ (I)
136	Arginine	Cysteine	✓ (0.45)	✓ (II)
461	Alanine	Threonine	✓ (0.23)	-
227	Arginine	Glutamine	✓ (0.15)	✓ (III)
68	Valine	Methionine	✓ (0.07)	✓ (III)
126	Asparagine	Aspartate	SNP (0.5)	-

3.3 Methodology: all-atom

The starting structure for all the simulations on the wild type was the human deletion mutant of G6PD structure (PDB code 2bhl), in which the first 25 and the last 9 residues have been removed [81]. All the mutant structures were obtained by mutating the wild-type using mutmodel [102], a program that allows the replacement of a single side-chain using the Minimum Perturbation Protocol (MPP [103]). Once a residue is specified, the side-chain is replaced and the Chi angles are rotated to minimise bad contacts until the best position is found. The obtained structures were checked for correctness and were fed as input into the family of tools of the GROMACS package [104–107] for the MD simulations steps.
All the simulations were performed using the AMBER99SB-ILDN force field [108], combined with the TIP3P[109] water model and using both versions 4.6 and 5.0.4 of GROMACS. AMBER99SB-ILDN is an improved variant of AMBER-99, containing backbone and torsion parameters that better fit NMR data. Other force fields (CHARMM [84] and GROMOS [110]) were considered, but eventually abandoned in favour of the more used AMBER force field. The main reasons behind this decision were that:

- GROMOS is a united atom force field, meaning that it is much less accurate than AMBER. The use of GROMOS is legitimate when computational time is taken into account as a simulation with GROMOS is generally faster than both AMBER and CHARMM. However simulations have proven that the time gain is not sufficiently different to justify the use of GROMOS.
- One of the main differences between AMBER and CHARMM is that CHARMM severely over-stabilises the formation of helical structures while AMBER tends to underestimate their stability, but overall both AMBER and CHARMM perform equally well. Because of the version used, AMBER matches NMR experimental data slightly more closely, especially at high temperatures [111, 112].

3.3.1 GROMACS topology and box creation

The first step was the creation of GROMACS coordinates and topology files using **pdb2gmx**. This tool takes a structure (e.g. a pdb file) and generates three outputs: a GROMACS-format coordinate file (.gro), a force field-compliant topology file (.top) and a position restraint file (.itp). The next steps consist of building a triclinic box around

the protein (**editconf**), of solvating (**genbox**) and adding counter-ions to the system (**grompp** and **genion**). The simulation box is a critical factor in every MD experiment; the bigger the box is, the larger the amount of atoms considered for the energy calculations (greatly affecting the performance). Since MD experiments use periodic boundary conditions, a minimum image convention must be set, this is because a protein must never interact with its periodic image, otherwise the forces calculated would be incorrect. For the first simulations at 310 and 500 Kelvin, the distance to the edge of the box was set to 1.5 nm, but because the first analyses indicated a stable system this distance was reduced for the following simulations, First to 1.4 nm first and then to 1.3 nm. The smaller boxes allowed an increase in performance, going from around 9 ns/day to 10 or 11 ns/day. Table 3.2 lists the boxes used in the simulations.

protein-edge distance [nm]	dimensions [Å]	water molecules
1.5	119.762 x 144.124 x 113.207	59347
1.4	117.762 x 142.124 x 111.207	56957
1.3	115.762 x 140.124 x 109.207	53972

TABLE 3.2: Approximate dimension and number of water molecules of the water boxes used in the simulations.

3.3.2 Energy minimisation

Before the beginning of the real dynamics, clashes or unfavourable geometries must be removed from the system by performing Energy Minimisation (EM). A maximum of 50000 steps of steepest descent were called and each event was considered successful when the maximum force on any atoms was less than 10 kJ/mol/nm. Some of the other important parameters used included:

- particle-mesh Ewald (PME) for the treatment of long-range electrostatics [113];
- twin-range cut-off for the van der Waals interactions;
- cut-off distance of 1 nm. The same value was used for the short-range neighbour list, long range electrostatic and long range van der Waals interactions.
- Periodic boundary conditions (pbc) along x,y and z directions (xyz).

These parameters were chosen mainly because they are compliant with the way the AMBER force field was parametrised.

3.3.3 System equilibration

Following the EM, solvent and ions must be placed uniformly around the protein; this is done in two steps: Initially the system was brought to the desired temperature in a 100 ps NVT canonical ensemble dynamics (NVT is the ensemble in which the number of atoms, the volume and the temperature are kept constant) and then pressure was applied to the system until it reached the density of 1 bar, using an NPT isothermal-isobaric ensemble (contrary to the NVT ensemble, in the NPT ensemble the pressure is kept constant, together with the temperature and the number of atoms in the system) for another 100 ps. During both phases the protein is frozen (-DPOSRES flag) to maintain the geometrical conformation obtained in Section 3.3.2. Important parameters for the equilibration phases included:

- Initial velocities were randomly generated (gen_vel=yes) using different seeds (gen_seed) for every simulation;
- The Berendsen thermostat was used for coupling the temperature [114]. As described in the introduction (Section 2.3.1), the temperature of a system is controlled by rescaling the velocities at each step, by multiplying the current temperature by a factor λ. However, in this way, the kinetic energy maintains constant values over time and the correct temperature fluctuations are not reproduced. The Berendsen thermostat overcomes this inaccuracy by coupling the system to an external bath only at certain time steps.
- The Parrinello-Rahman barostat was used for coupling the pressure [115]. Similarly to the Berendsen thermostat, the Parrinello-Rahman barostat couples the pressure to an external pressure bath, allowing both the volume and the shape of the simulation box to fluctuate.

3.3.4 Production MD

The parameters were almost the same as the NPT step, with the only exception that the Berendsen thermostat was replaced by the Nose-Hoover thermostat [116, 117] to produce a better kinetic ensemble. The Berendsen thermostat tends to converge quickly, and is thus more suitable for equilibration. Once the system is near temperature convergence, the Nose-Hoover thermostat kicks in to produce the correct energy fluctuation. Other parameters used in the simulations included:

- particle-mesh Ewald (PME) treatment of long-range electrostatics [113]. PME works by converting the system into a grid of density values and calculating the forces applied to each particle depending on its position relative to the closest cell;
- An all-bonds LINear Constraint Solver (LINCS) algorithm [118] was used for handling constraints;
- Because of the presence of LINCS, the time step was increased to 2 fs. In MD simulations, the bond oscillation is one of the factors that limits the time step. By replacing the bond vibrations with constraints, it is possible to increase the time step. Algorithms like SHAKE [119] reset bonds to defined values one bond at a time, increasing accuracy, but it is too slow to be used on large systems. The LINCS algorithm uses Lagrange multipliers to model the constraint forces, resulting in a method that is three to four time faster than SHAKE.
- To be able to take advantage of the GPU acceleration, the Verlet cut-off scheme was used with an nstlist (frequency of updating the neighbour list) equal to 30. The Verlet scheme keeps a list of particles within a certain cut-off distance;
- The conformations were saved every 10 ps, with the exception of the first wild-type replica (310 K) for which the structures were saved every 20 ps.

All the simulations ran on a single EMERALD [120] node, consisting on two 6-core X5650 Intel Xeons and three or eight 512-core M2090 NVIDIA GPUs. Initially the target trajectory length was 200 ns, but following the first results, this number was raised to 500 ns.

3.3.5 Analyses

All the trajectories have been analysed using an in-house script (doitGROMACS.sh see Appendix A) and the statistical package R [121, 122] for plotting. The script is capable of automating the set up of an MD simulation and performing some standard analyses on the trajectories. Before analysing the trajectories, water was removed from the reduced-precision trajectories (.xtc); This step was not strictly required, but because of the dimension of the trajectories, it proved much more convenient to remove everything but the protein. The following analyses were carried out:

- The root mean square deviation (**rmsd**) was calculated on the protein backbone using *g_rms*;

- The **gyration radius** was calculated on the entire protein using *g_gyrate*;
- The root mean square fluctuation (**rmsf**) was calculated for both the backbone and the side-chains using *g_rmsf*;
- **Cluster** analyses on the backbone, were performed using *g_cluster* on a rmsd matrix generated by *g_rms*;
- Principal Component Analyses (**PCA**) were performed on the protein C-alphas using *g_anaeig* on a covariance matrix generated by *g_covar*. *G_sham* was used to obtain a profile of the Potential Energy Surface (PES) from the first eigenvectors.
- Solvent Accessible Surface (**SAS**) analyses on the entire protein were performed using *g_sas*. The SAS algorithm builds the area that is accessible to the solvent by rolling a sphere of defined size (a probe) over the Van der Waals surface of the protein. Because we were interested in seeing the areas of the protein which are accessible to an antibody, a probe bigger than the default size was used. From existing literature, it is known that the average area of interaction is 15 $Å^2$ [123], here a probe size of 0.7 nm in diameter and 24 dots per probe. The probe size was reset to the default value of 0.14 nm when only the binding sites (Glucose, co-enzyme and structural $NADPH^+$) were considered. This was done to allow the detection of small changes in surfaces values.
- Secondary structure analysis (**dssp**) on the main chain was performed using *do_dssp*.
- The hydrogen bonding count (**hb**) was performed on the entire protein as well as all the binding sites using *g_hbond*.

  Refer to Appendix A for a full explanation of the script and the parameters used.

3.4 Wild-type characterisation

Before analysing the damaging effects of the mutants, and better to highlight the differences between them, it is important to characterise the wild-type and its behaviour.

3.4.1 Analysis with Elastic Network Model (ENM)

Normal Mode Analysis (NMA) is a computational technique used to study large scale motions in biological molecules. NMA is based on the assumption that normal modes with the largest fluctuation (low frequency) are the motions which have biological relevance. NMA uses a harmonic potential (Equation 3.1) to describe the fluctuation around a minimum energy [124]. To perform NMA, the energy of the system is first minimised, then the Hessian matrix is calculated and the eigenvalues and eigenvectors are finally extracted by diagonalising the Hessian matrix. Because these steps are particularly computation expensive, several simplifications have been proposed over the years. Elastic Network Models (ENMs) are NMAs in which the system is dramatically simplified into an elastic mass of atoms (Cα) connected by springs.

$$E_p = \sum_{d_{ij}^0 < R_c} c(d_{ij} - d_{ij}^0)^2 \tag{3.1}$$

were d_{ij} is the distance between two atoms, c is the potential spring constant, and R_c is an arbitrary cut-off, beyond which interactions are not considered. The ENM implementation used in this project was the "*elNémo web-interface to the Elastic Network Model*" [125, 126] and was mainly used to locate the region of G6PD that has a normal propensity to motion. The elNémo web-interface is easy to use and only needs a structure (PDB file) to run. All the basic settings for the calculation were left at their default values. The B-factor extracted from the X-ray crystallography structure (2bhl.pdb), indicates that the extremities are the most flexible regions of G6PD (Figure 3.3). There is a noticeable difference between the two chains of the protein, and this could be connected with spatial inhomogeneities in packing density, that could cause variations in small-amplitude structural flexibility within the protein [127]. High-density regions can accommodate only a few similar conformations, while low-density regions might allow several conformations. With more blue shades, the N-terminus of chain B (right part of Figure 3.3) seems to fluctuate less than that of chain A, and this difference is swapped in the C-terminus, where chain A is more stable. Apart from the extremities, other groups of residues (mainly around loops and short spanning helices) present higher B-factors.

In particular, the regions from residues 311 to 330, 240 to 250 and around residues 423-428. The B-factor predicted by elNémo (Figure 3.2) over-estimates the flexibility at the N-terminus, but it overall coincides with the crystal structure values. Both sources indicate that the β core is the most stable region of the protein.

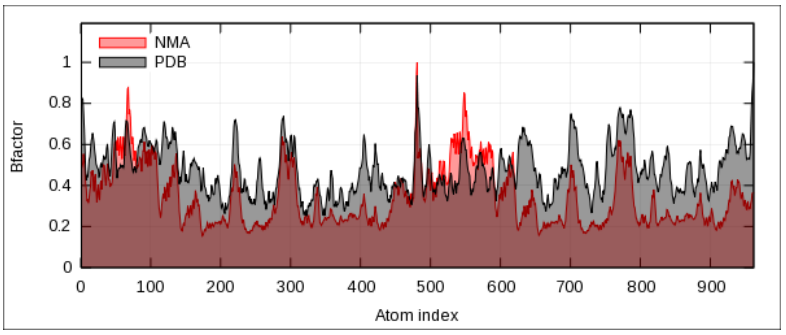


FIGURE 3.2: Comparison between the B-factor predicted by the NMA method and the data in the PDB file (2bhl). The first 500 residues constitute the first chain, while the next 500 are from the second chain.

ENM was also used better to understand the global movements of G6PD, and thus what we may have expected from the simulations. The natural movements of G6PD are represented by arrows in Figure 3.4, and in both figures, the N and C-terminus close and twist around the central β domain. These movements may favour the uptake of the substrate by allowing or denying access to the binding site.

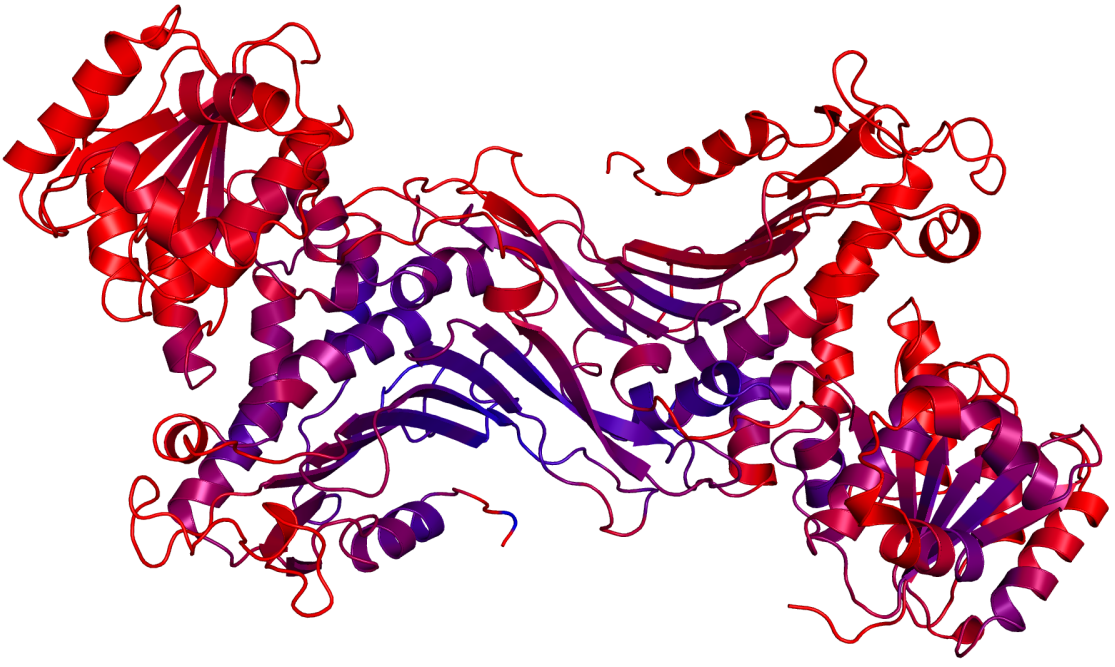
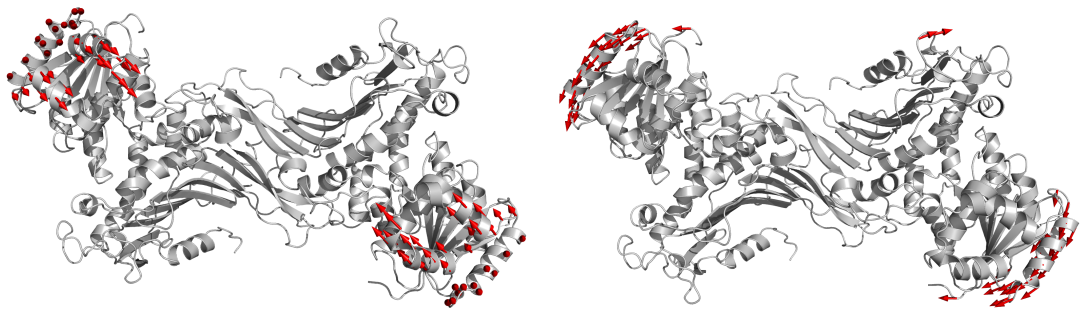


FIGURE 3.3: G6PD coloured following the B-factor values included in 2bhl.pdb; from blue (low fluctuation) to red (high fluctuation).



(a) (b)

FIGURE 3.4: The figure shows the first two non-trivial modes obtained from the elNémo calculations. The arrows are used to follow the conformational change that are induced by the modes.

3.4.2 Wild-type at 310 K (37°C)

The wild-type structure (2bhl.pdb) was simulated at 310 K (approximately 37°C) with three independent simulations (replicas). The initial coordinates were shared among all the trajectories, while the velocities were different. This was done to sample the conformational space better. Two replicas ran for 250 ns, while a third continued only for 200 ns.

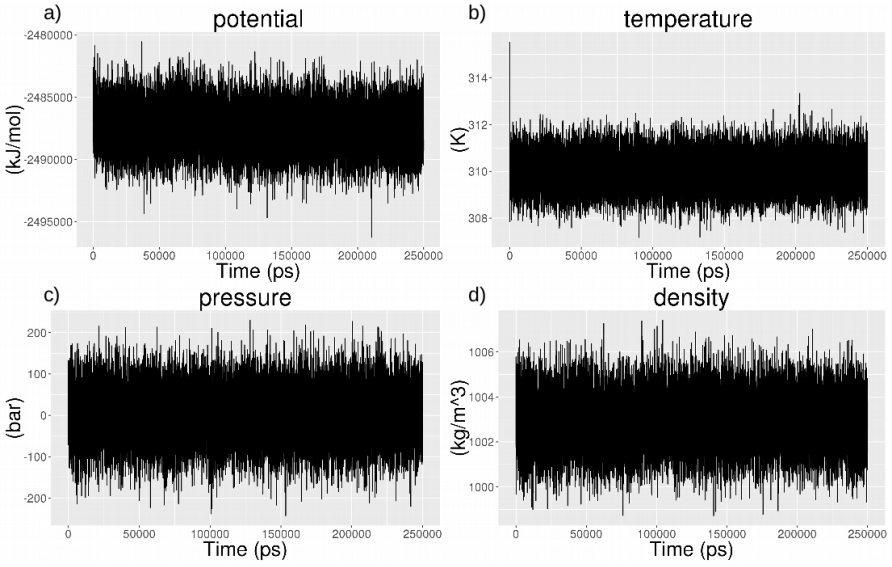


FIGURE 3.5: The stable fluctuation of (a) the potential energy, (b) the temperature , (c) the pressure and (d) the density of the simulation box during a dynamics of the wild-type at 310 K.

The simulation conditions (Figure 3.5) were well preserved in all the replicas, meaning that the system is stable and that the simulation parameters were set correctly. The potential energies are in the order of -2.49 $\times 10^6$ kJ/mol with the system density oscillating around 1002.9 kg/m^3. Even though each of the replicas presented differences in their

dynamics, all of them were similarly stable, with a rmsd (root mean square deviation) within a 0.3 nm range. Only one replica (marked in blue in Figure 3.6) contained peaks exceeding 0.3 nm towards the end, but the profile is not dissimilar to the others.

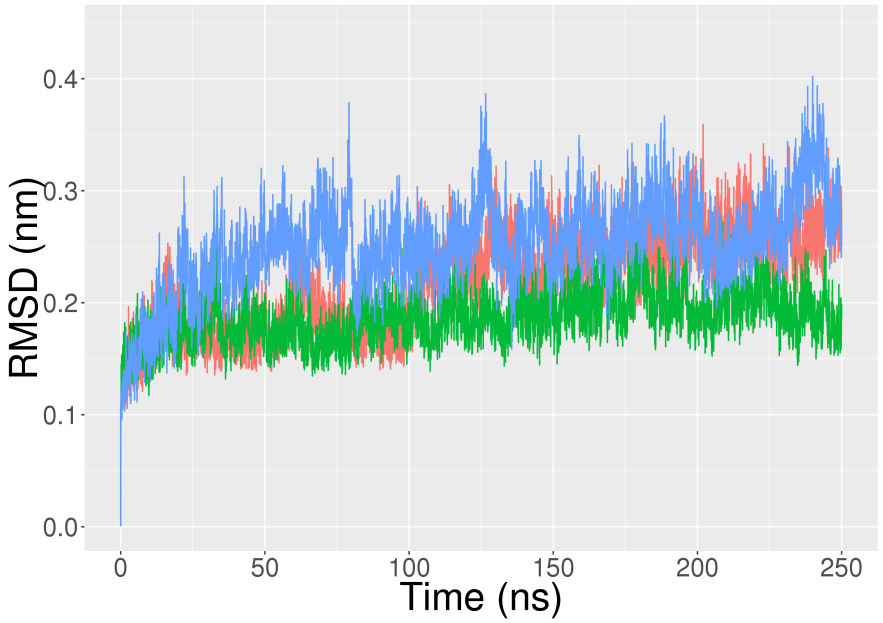


FIGURE 3.6: Comparison of rmsd for the three replicas of the wild-type at 310 K: replica 1 is in red, replica 2 in green and replica 3 in blue.

An observation of the trajectories reveals that all the domains are stable and only the loop regions are active in a constant rearrangement. The protein maintains its compactness with only the external $\alpha+\beta$ 3-layer(aba) sandwich domains slightly oscillating on their axes. This is made visible also by the radius of gyration which oscillates between 3.6 nm (which is the value calculated for the initial PDB structure used) and 3.7 nm. The movements observed at this stage are very similar to the ones outlined with elNémo (Section 3.4.1) and are described by a mechanism of opening and closing. The overall stability is confirmed also by the root mean square fluctuation (rmsf) analysis (Figure 3.7), in which it is possible to have an indication of the single fluctuation of the residues over time.

In all the replicas, the residue fluctuations are generally below 0.2 nm, with very few peaks exceeding this value. Analysing these profiles, it was possible to locate a pattern of fluctuation, that is seen and maintained in almost all the simulations of both wild-type and mutants. This pattern consists of two groups of residues with rmsf values above the average. The first group includes the N-terminus region between residue A71, for which the rmsf begins to grow and residue Q133, for which the rmsf drops to zero. The second group includes several other residues, scattered along the entire sequence, for which the rmsf has peaks as high as, or higher than, the N-terminal region described before

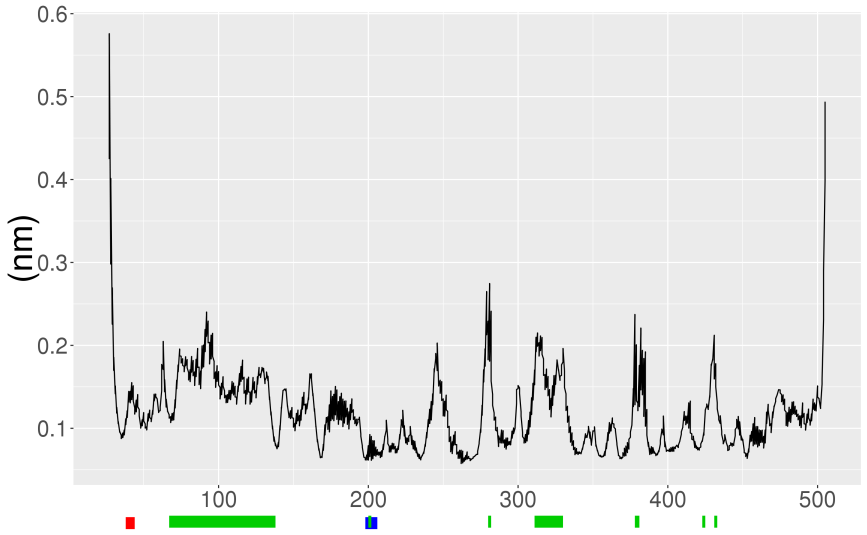


FIGURE 3.7: Rmsf profile of one of the replicas of the wild-type at 310 K. At the bottom, the binding sites are coloured in red and blue (respectively co-enzyme and substrate), while the regions of higher fluctuation are in green.

(A71-Q133). These residues are: R175, R246, D282, the region spanning from N311 to P329, I380 and K432 (Figure 3.8), and are highlighted in green below the rmsf graphs (e.g. Figure 3.7). It is particularly interesting to note that the N311-P329 portion is home to a short three-residue helix which links the central β domain to the α helix of the α+β 2-layer sandwich domain that constitutes the back of the dimerisation interface (Figure 3.8 in red).

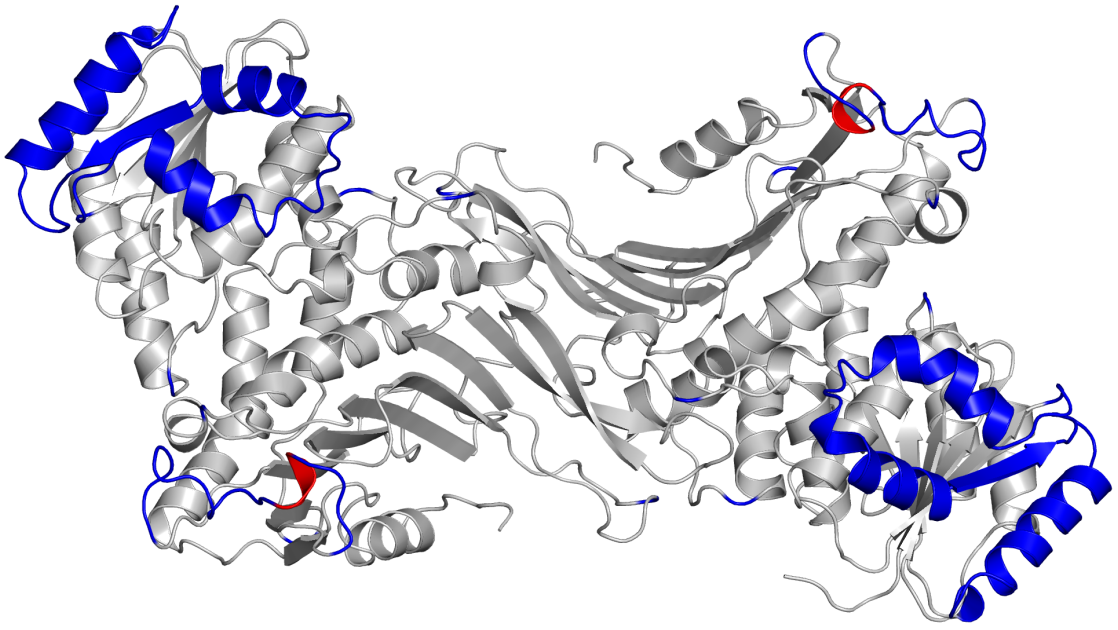


FIGURE 3.8: The residues that present the highest fluctuations are represented in blue. The helix delimited by residues N311 and P329 is coloured in red.

Principal Components Analysis (PCA) helped determine the motions which contribute the most to the dynamics of the protein. All the replicas present movements compatible with the ones already detected by elNémo (Section 3.4.1), in particular the opening and closing over the central β hinge of the two external globular domains (Figure 3.4), which was described earlier. This mechanism may permit or deny access to the binding site; when the domains are close together the glucose can neither bind to, nor escape from, the active site, while a relaxation of the structure may allow its uptake. With the PCA, it is also possible to combine the eigenvectors that account for most of the movements, in order to obtain a profile of the potential energy surface (PES) explored during the simulation (Figure 3.9). The dynamics were able to sample more than one point of minimum energy, and a comparison between the PES and the rmsd profiles can account for the difference in rmsd seen in the three replicas. The more energetic replicas explored two clearly distinct points of minimum energy, while in the other replicas, there was not enough energy to climb the saddle point between the two minima. A sudden change in rmsd, seen after the first 100 ns, may correspond to the moment in which the system shifted from a lower energy structure to a slightly more energetic conformation.

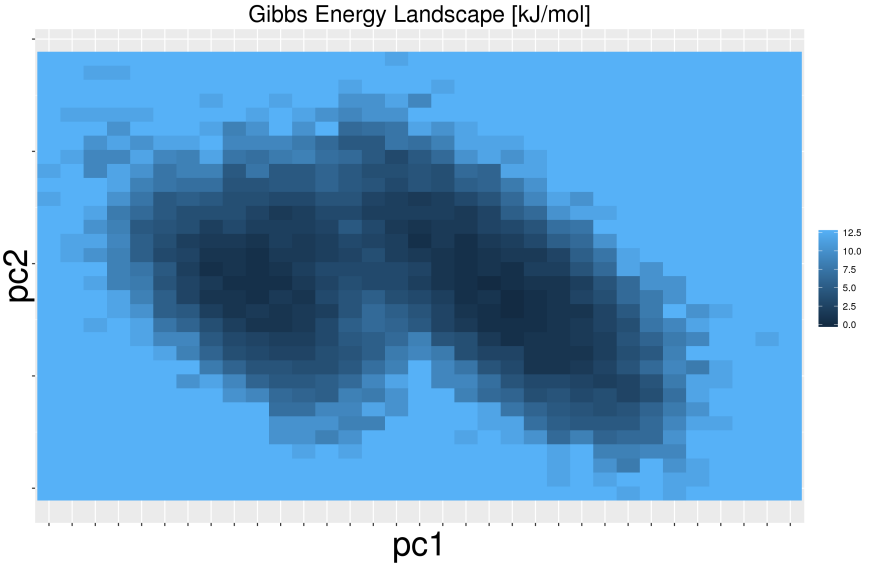


FIGURE 3.9: The first two eigenvectors are projected together to define the PES of one of the most energetic dynamics. The dynamics overall explored two well defined minimum.

Of a total of 960 residues, more than 700 retain their initial secondary structure during the simulations. The α helices are the most abundant, with β-bridges (longer hydrogen bonds) and β-sheets following. It is crucial that, at this stage, there is no interconversion between secondary structures, i.e. a residue in a helix stays in a helix.

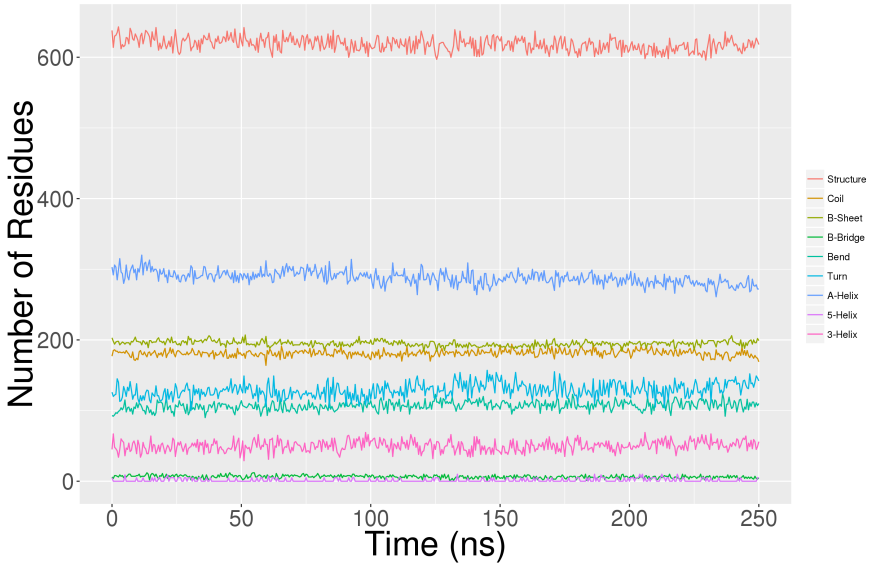


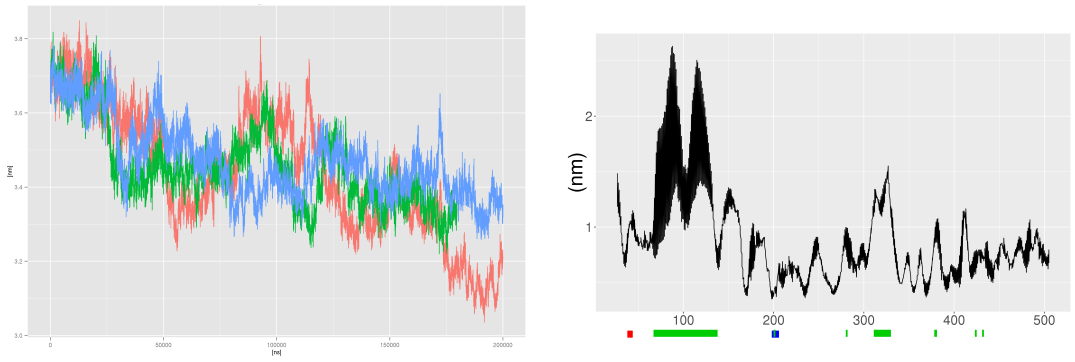
FIGURE 3.10: The count of the number of residues for each secondary structure type of the first replica indicates that the residues maintained their assigned secondary structure for the entire simulation.

3.4.3 Raising the temperature: Wild-type at 500 K (226°C)

The goal of an MD experiment is to explore the PES of a molecule to look for structures which have low values of potential energy and are therefore considered stable. At low temperature, it may happen that the dynamics gets stuck around one of these points and keeps sampling the same region. To overcome this limitation and to extend the sampling, it is common practice to increase the temperature to boost the energy of the system. The dynamics will then have enough energy to overcome saddle points and escape from points of minimum energy.

The G6PD wild-type was studied at 500K with three different replicas: two of 200 ns and one of 179 ns. The latter simulation failed before reaching 200 ns and is representative of the great instability in such conditions. At this temperature, it is difficult to distinguish random motions from important ones. In fact the rmsd comparison (Figure 3.11) shows how the structure changes constantly from the beginning to the end of the simulation, reaching a probable plateau near the end of the simulations, probably owing to the collapse of the structure on its own centre of mass.
Correlated with the rmsd, the radius of gyration dramatically decreases to reach a value of 3 nm (Figure 3.12). Contrary to the previous simulations, here the rmsf profiles jump to extreme values of over 1.5 nm, as a result of the Rossmann-like domain and the regions of higher fluctuations being highly exposed.
At high temperature, the number of residues which retain a defined structure drops

FIGURE 3.11: Comparison of the constantly growing rmsd for the three replicas at 500 K: Replica 1 is in red, replica 2 in green and replica 3 in blue.

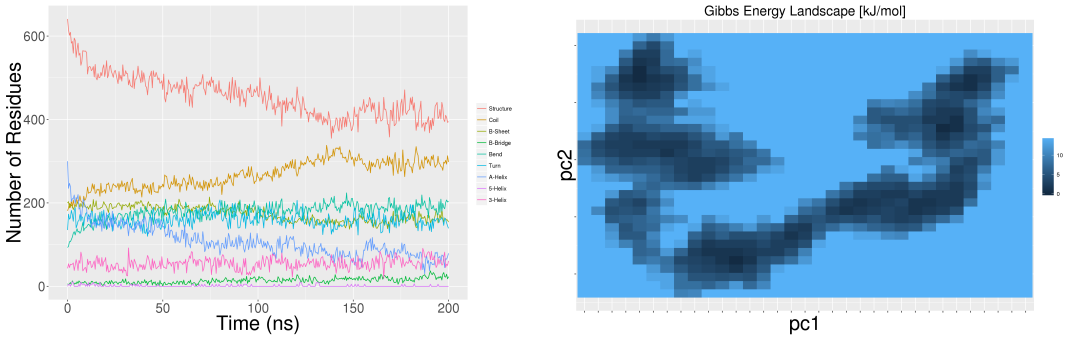


(a) (b)

FIGURE 3.12: (a) The radius of gyration of all the replicas drops dramatically during the dynamics. (b) The great fluctuations of the residues at the N-terminus observable through the rmsf profile of one replica of the wild-type at 500 K.

from 600 to 400 (Figure 3.13). A deeper analysis suggests that most of the lost structure comes from conversion to coils mostly from α helices and only partially from other structures. There is a slight decrease in β sheet composition, but these regions tend to be stable for the whole simulation, maybe because the inward collapse of the structure protects these regions from unfolding.

The α-coil conversion is confirmed by the visual inspection of the trajectories that indicate that the long and stable β sheet constitutes the central axis for the movements. The N-terminus tends to move inwards, toward the centre of the protein, resulting in a very compact final structure, where access to both the G6P and Co-enzyme sites is obstructed (Figure 3.14). When the first two principal components are plotted together (Figure 3.13b), it appears clear that the system is far from reaching convergence. These simulations were important because they suggested that, if an unfolding event occurs,



(a) (b)

FIGURE 3.13: (a) Secondary structure evolution of one of the replicas at 500 K. (b) PES profile, obtained by combination of the first two principal components.

it may cause the collapse of the G6PD structure rather than its explosion. With this in mind, the simulation box in the following simulations was reduced in size. A smaller box brings a slight increase in performance owing to a smaller number of solvent atoms in the system. At the end of the analyses, it was clear that 500 K was too high for G6PD to maintain its structure, and it was decided to try a temperature that was between 310 and 500 K. The hope was to find a temperature low enough to keep the system stable, but high enough to see important structural modifications resulting from the mutation clearly. To avoid the risk of choosing another temperature that would prove to be too high, it was decided to stay closer to 310 K and 400 K was selected.

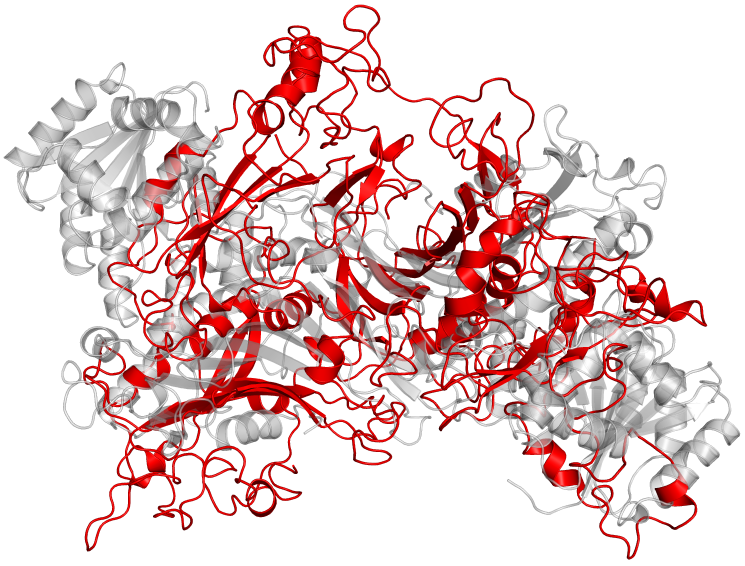
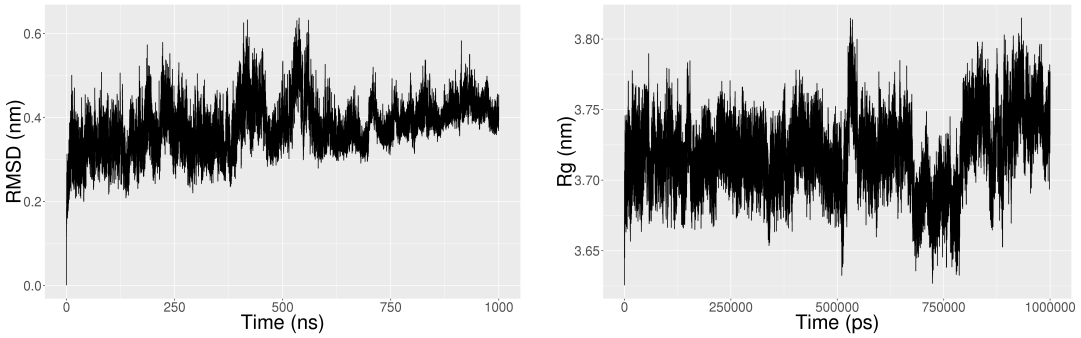


FIGURE 3.14: Misfolded structure of the wild-type (red) compared with the folded one (gray).

3.4.4 Wild-type at 400 K (126°C)

Initially two 200 ns long replicas were performed, one of which was extended to 1 μs (1000 ns). This section will mainly describe the behaviour of the longer replica, describing the second one only when the two behaviours greatly diverge. Similarly to what happened at lower temperature, the simulation conditions did not present unusual fluctuations, indicating that the simulation was stable. As expected, the potential energy oscillates around -1.99 $\times 10^6$ kJ/mol and the system density is around 906 kg/m^3, respectively higher and lower than the wild-type at 310 K and 500 K (Table C.1). The protein looks stable in its rmsd, drifting only 0.4 nm from the initial conformation (Figure 3.15a). The rmsd starts at 0.3 nm and slowly reaches 0.4 nm by the end of the simulation with only two brief peaks over 0.5 ns after 400 and 500 ns respectively.



(a) (b)

FIGURE 3.15: During the 1μs trajectory at 400 K , both (a) rmsd and (b) radius of gyration indicate that there are no changes in fold for the wild-type.

For most of the simulation, the radius of gyration oscillates around a constant value of 3.72/3.73 nm. 670 ns into the simulation, the radius of gyration suddenly drops to 3.65 nm for 110 ns. The protein rapidly regains its fold and maintains it for the last 100 ns (Figure 3.15b). In Figure 3.16, the structures taken at different moments of the trajectory are superimposed. From the figure it is possible to see how the Rossmann-like domain bends slightly towards the centre with a movement similar to that described in Sections 3.4.1 and 3.4.2. In addition, the unstructured residues from A300 to R330 are free to move, affecting the radius of gyration calculation.

The rmsf does not differ very much from the ones previously described, and only few residues had peaks above 0.25 nm. Overall the simulations show a stable first half of simulation and a more dynamic second half.

Even though the tendency of the α helices to unfold in favour of coil is present (Figure 3.17), the conformational changes are only 0.3 nm, suggesting that differences are

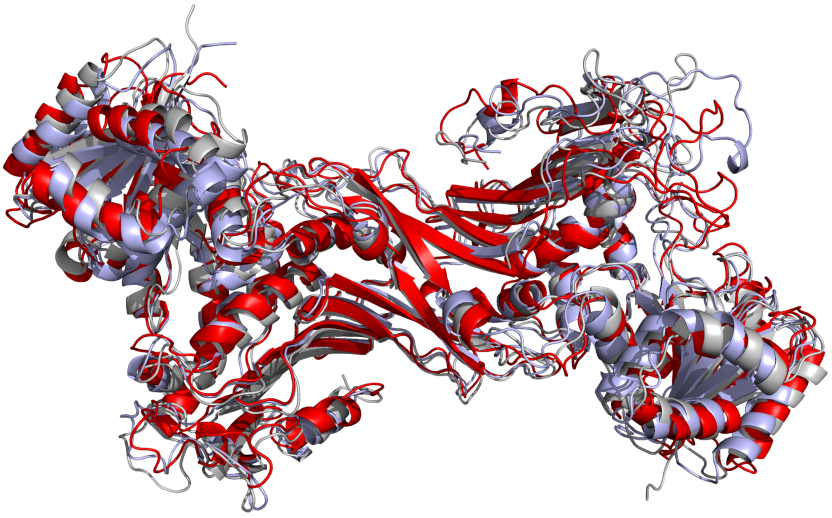


FIGURE 3.16: Superimposed structures at different moments of the trajectory, from beginning (grey) to end (red). The structure corresponding to low radius of gyration (670 ns) is coloured in light blue.

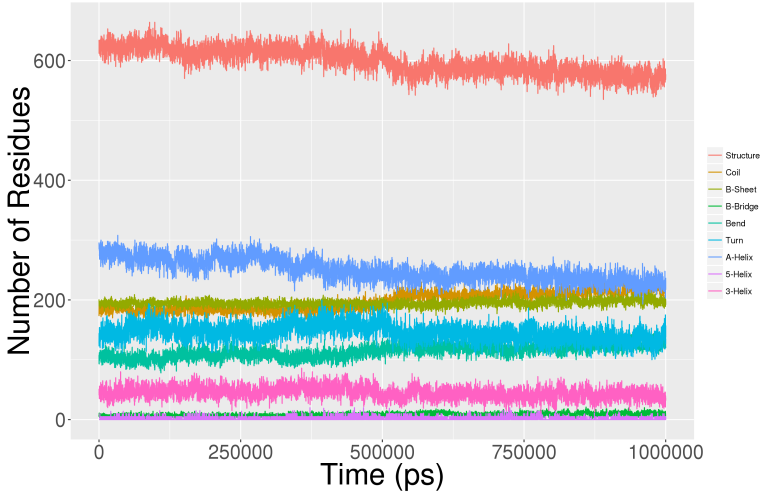


FIGURE 3.17: The total number of residues that has a defined structure drops during the simulation, as a result of helices converting into coils.

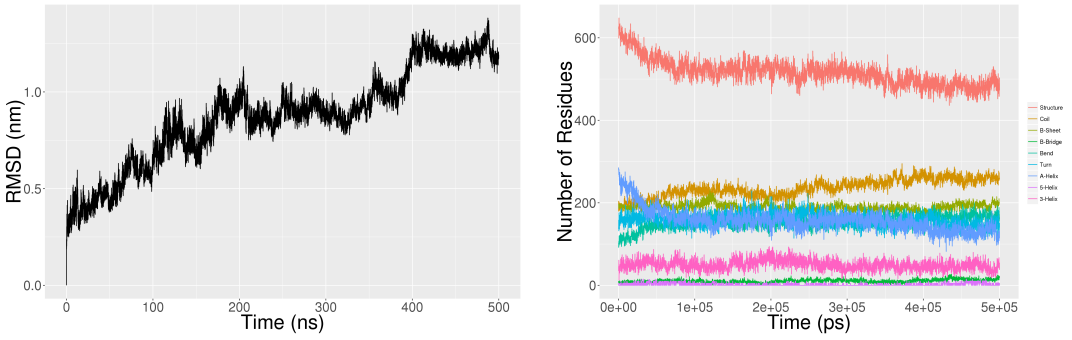
mainly caused by the movements of the loops (Figure 3.16). Solvent Accessible Surface area (SAS) analyses were performed on both the entire structure and the inside of the binding sites (glucose, co-enzyme and structural NADPH+), in order to detect small changes that are not able to disrupt the fold, but that could be strong enough to affect the substrate binding. At 400 K, the protein presented a SAS area that was similar in size to that found at room temperature (Table C.2). The total area (calculated with a probe of 0.7 nm in radius, as explained in Section 3.3.5) was only 7 nm^2 bigger, while the difference was reduced to only 2 nm^2 when the default probe, 0.14 nm in radius, was used. This allowed the calculation of the area inside the binding site, which was exposed to the solvent. The glucose binding sites had an average surface of 8.48 nm^2, the

Co-enzyme site of 14.47 nm^2 and the structural $NADP^+$ site of 12.82 nm^2 (Table C.2).

The data collected so far indicated that 400 K is a temperature at which G6PD is still stable, but it is also high enough to start detecting some signs of instability and behaviour change. This suggests that 400 K could be the target temperature at which it should be possible to show the destabilising effects of the mutants. To be sure of having found the right temperature, a simulation where the temperature was raised by 50 K was performed.

3.4.5 Wild-type at 450 K (176°C)

When the temperature is raised by 50 K, the protein presents a behaviour very much similar to the dynamics at higher temperature (500 K). The system density is already 200 kg/m^3 lower dense than at 310 K and the potential energy has risen to -1.92 $\times 10^6$ kJ/mol (Table C.2). Unlike to the β regions that are well defined, the α-helices of the N-terminus of the Rossmann-like domain have already started to unfold during the equilibration period. The rmsd reaches a plateau around 1.3 nm (Figure 3.18a), further proof of the instability of the protein in these conditions.

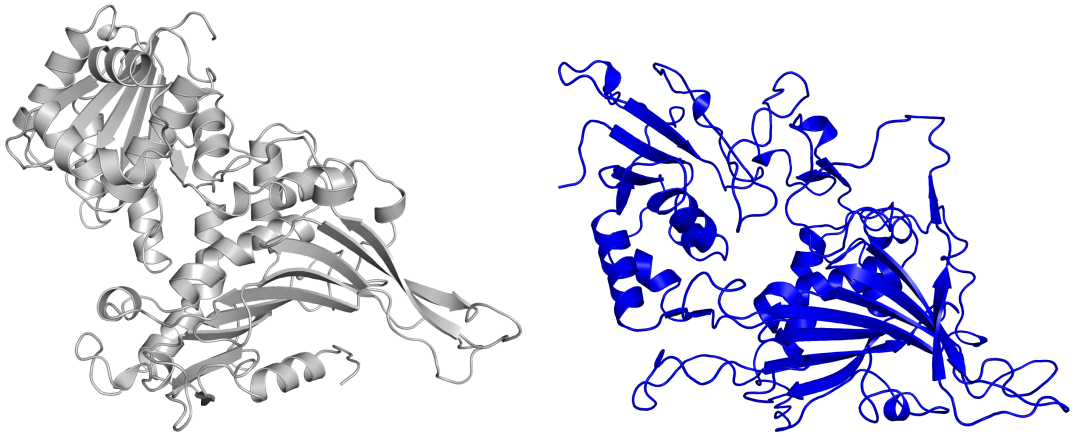


(a) (b)

FIGURE 3.18: (a) Rmsd of the simulation at 450 K. (b) Secondary structure count of the simulation at 450 K.

The central β-sheet region is always clearly visible, but it moved to a more compact configuration, while the percentage of α helices drops by 50% (Figure 3.18b), with both N-termini almost completely unfolded close to the end of the simulations (Figure 3.19).

The dynamics results in a structure that is flattened and more extended laterally compared with the wild-type. In this configuration all the binding geometries are lost and it is unlikely that the co-enzyme would bind and interact with the substrate. After the simulation at 450 K, it was clear that also this temperature was too high for maintaining



(a) (b)

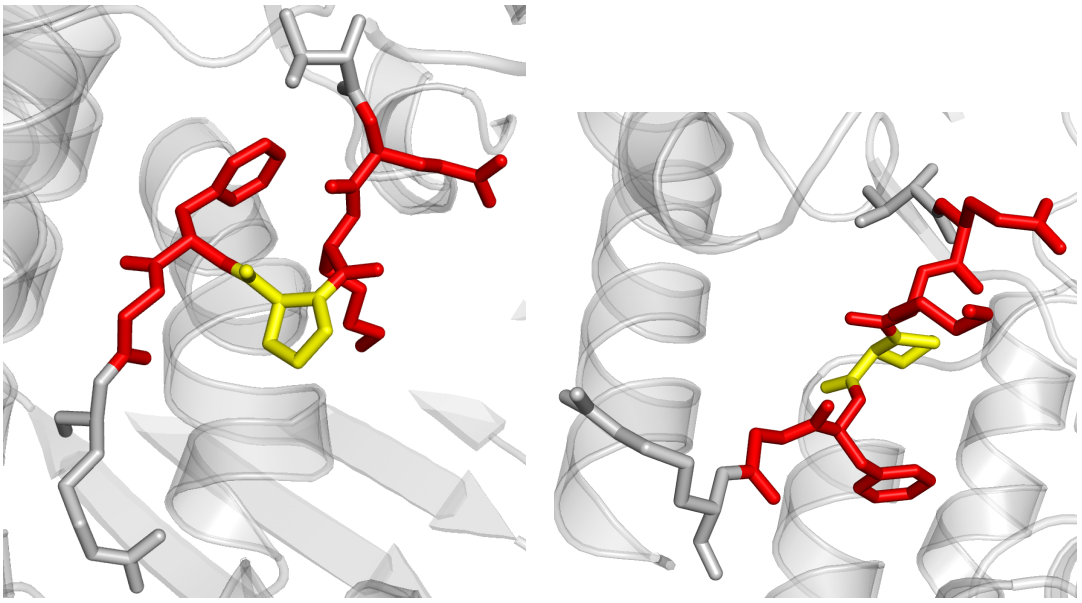
FIGURE 3.19: G6PD structure at the beginning (a) and at the end (b) of the simulation at 450 K.

G6PD stability, and 400 K was recognised as the closest value to the temperature that causes G6PD to lose its stability. 400 K was then taken as the reference temperature for all the future simulations.

3.4.6 Proline 172 cis-trans isomerisation

Proline 172 is the central residue of the conserved peptide EKPxG. This residue has a key role in G6PD, because it allows the correct positioning of both the substrate and the co-enzyme in their respective binding sites [81], by allowing Lys171 to interact with both G6P, through its terminal amino group, and $NADP^+$, through the carbonyl group. This function was proposed to be the result of its *cis-trans* isomerisation (Figure 3.20), which blocks the conserved EKP turn to restrict the access of the nicotinamide ring. Humans tend to exhibit the *trans* form, but both forms have been found in G6PD structures. In particular it seems that the *cis*-Pro172 has a more mobile co-enzyme binding site and thus is found in the absence of the binding with it. The only G6PD variant known to involve this residue (Volendam: P $\rightarrow$ S) exhibits class I phenotype and has the lowest K_m values than any other known variant [128].

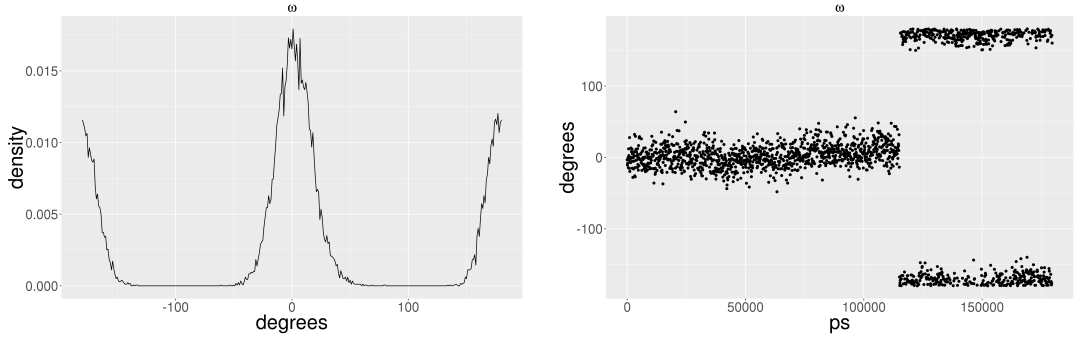
The *cis-trans* isomerisation could be a real feature of G6PD or simply an artefact of the crystal structure. In an attempt to answer this question, and to see the Pro172 isomerisation, the values of the omega (ω) angle of the peptide bond between the proline and its preceding residue (lysine) were monitored during the various dynamics simulations. As described in Section 3.3, the wild-type structure was taken from a complex (PDB file: 2bhl) without the co-enzyme, where the Pro172 adopted a *cis* conformation ($\omega = 2.5°$).



(a) (b)

FIGURE 3.20: Pro172 in both (a) *cis* and (b) *trans* configuration obtained from the 1qki PDB file. Pro172 is represented in yellow, and the other residues of the EKPxG peptide are coloured in red.

At low temperatures (310 and 400 K), the omega angle remains around 0°, with values that oscillate by ± 20°. Raising the temperature, it is possible to observe a gradual increase in this range, up to the point at which the omega angles switched to the typical values of a *trans* conformation (Figure 3.21).



(a) (b)

FIGURE 3.21: (a) The density distribution and (b) the values of the omega angle of Pro172 in chain B for the dynamics of 500 K show the presence of both *cis* (ω around 0°) and *trans* (ω around 180°) conformation.

These results seem to confirm that the *cis-trans* isomerisation may occur in G6PD, even if it was observed only at very high temperature (500 K). A study on peptide bond isomerization at high temperature simulations [129], showed that all existing force fields tend to allow spontaneous isomerisation of cis-trans conformations in temperatures

higher than a specific value. For the AMBER force field, the threshold is 700 K [129]. The fact that Pro172 isomerised at 500 K (200 K lower than the threshold), and that was the only proline in G6PD to do so, suggested that the *cis-trans* isomerisation of Pro172 may be a real feature in G6PD, and that this feature could be important for the correct functioning of G6PD.

3.4.7 Wild-type summary

Wild-type G6PD presents a structure in which the fold is destabilised only at high temperature (above 450 K). The central $\beta-\alpha$ structure constitutes the core of the protein working as a hinge, supporting the movement of the external domains. These movements allow the opening and closing of the enzyme, facilitating the binding and interaction with both the substrate and the co-enzyme. The movements of the domains are facilitated by the presence of loops around each structure. This stability only breaks down at very high temperature (500 K), and at an intermediate temperature (400 K) the enzyme seems to retain its structure and features. The simulations at high temperatures helped to identify the weakest part in the G6PD structure, the N-terminus of the Rossmann-like domain and the helices.

FIGURE 3.22: All the mutants studied are represented as red spheres on the G6PD structure. A^-, the only multi missense mutation (mutations with at least two bases substituted) studied, is coloured in blue.

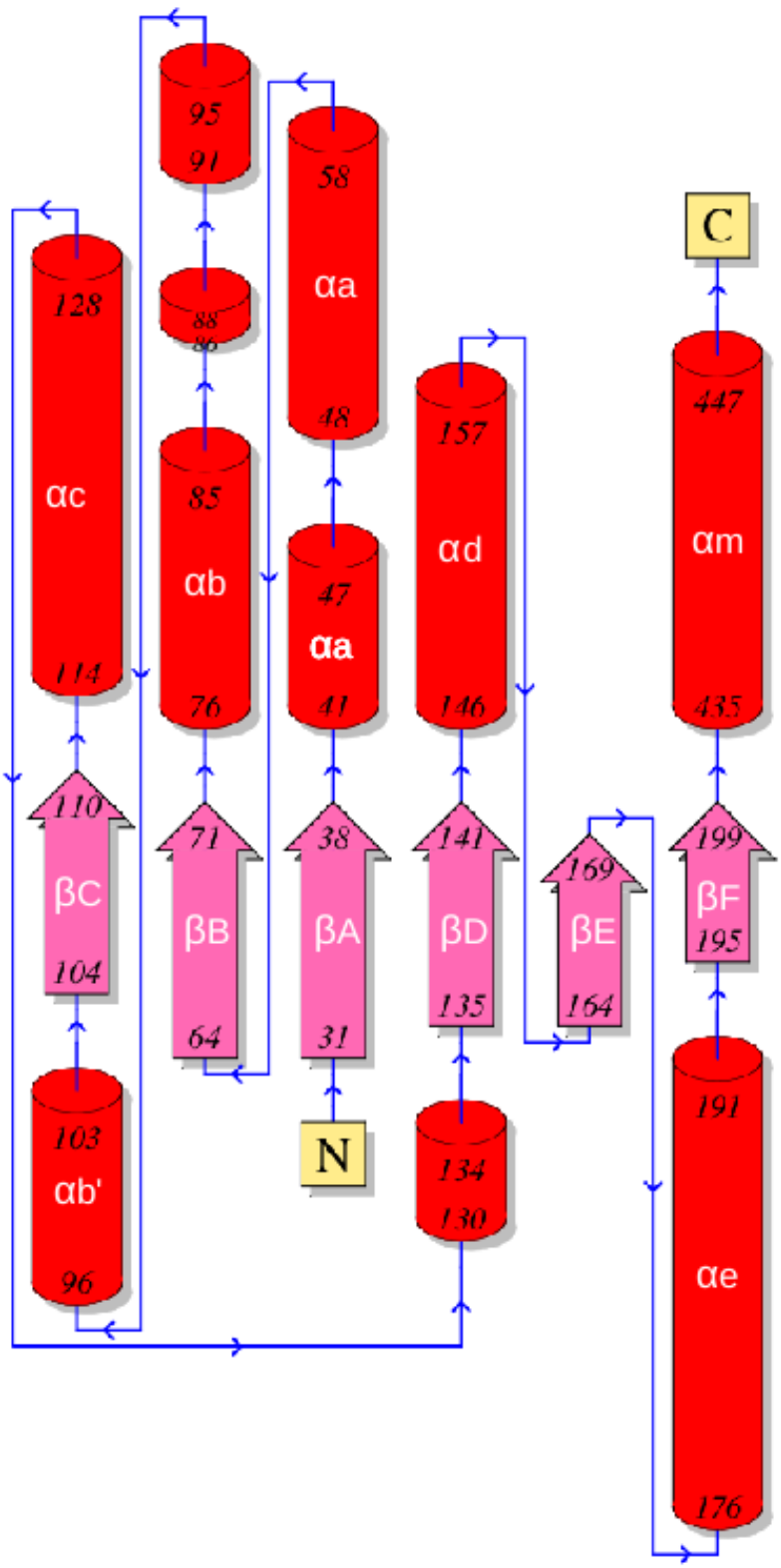


FIGURE 3.23: Diagram of the "NADP-binding Rossmann-like Domain" (CATH code: 3.40.50.720) obtained from PDBsum. The names of the helices and strands were added following the naming system used in the G6PD database of mutations [101]

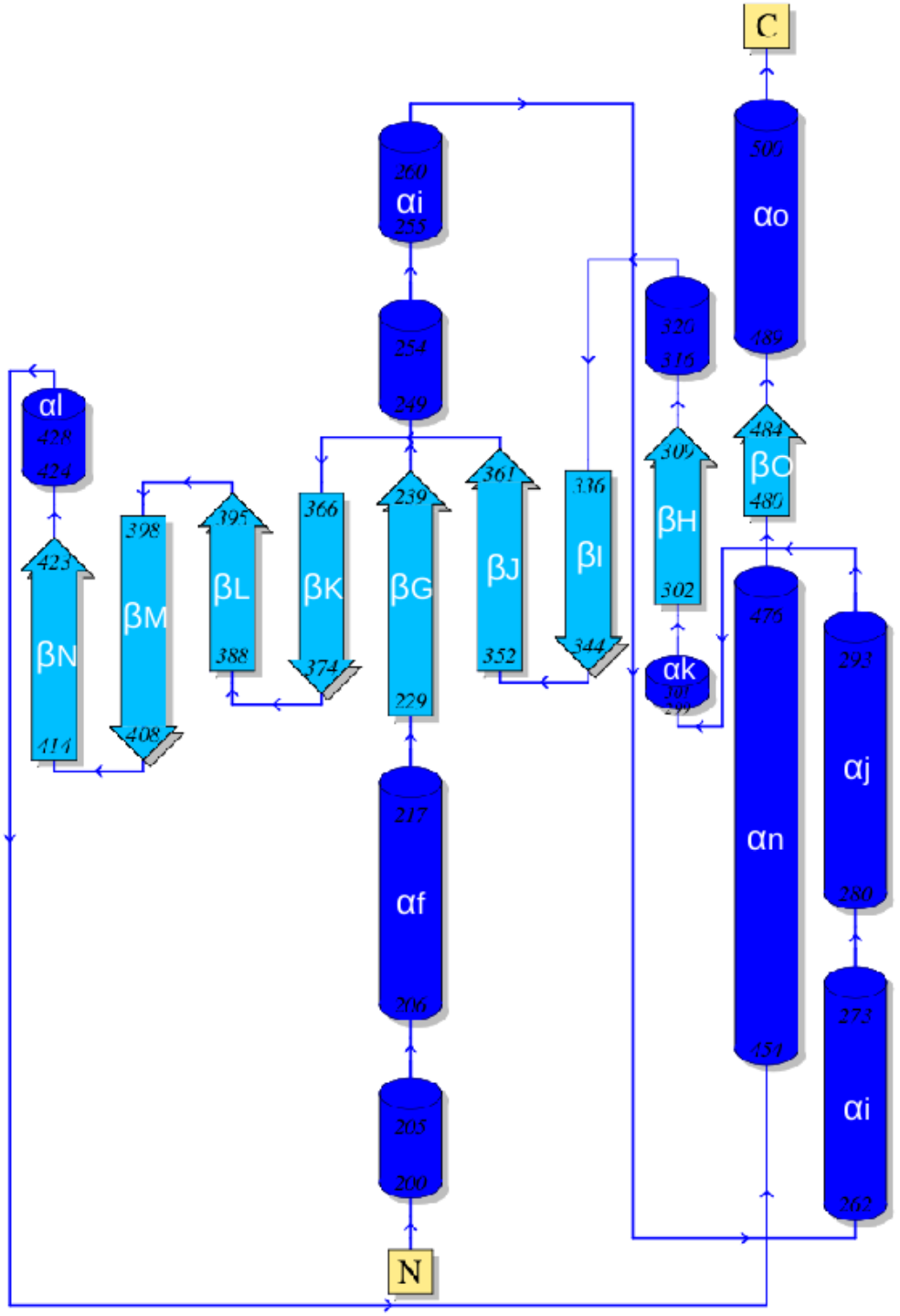


FIGURE 3.24: Diagram of the "Dihydrodipicolinate Reductase; domain 2" domain (CATH code: 3.30.360.10) obtained from PDBsum. The names of the helices and strands were added following the naming system used in the G6PD database of mutations [101]

3.5 Mutant simulations

A total of 17 G6PD mutants were studied (Table 3.1), and their location on the G6PD structure is shown in Figure 3.22. At the beginning of the project, for reproducibility and to sample the energetic profile of the proteins better, two replicas per mutant were started. Because the project had the aim of studying big conformational changes that might be caused by the mutations, and because of the high computational cost of the simulations, it was decided, for the last simulations and the mutants selected later in the project, to focus on a single and longer replica per mutant. As a result, some mutants have several shorter simulations while others have fewer and longer ones. To map the different structural features onto the G6PD structure, the α helices and the β strands were named using the lettering system represented in Figure 3.23 and Figure 3.24. Following the G6PD database of mutation [101], both the α helices and the β strands were ordered alphabetically from the N- to the C-terminus, using lower case letters for the helices (e.g. αf) and capital letters for the strands (e.g. βO). Some helices represented in the diagrams are not named, and this is because of some discrepancies between the representation given by PDBsum (from where the diagrams were taken) and the G6PD database, which grouped several small helices into a longer one. For consistency, the text follows the numbering scheme that was used in the G6PD database. Another convention found in the text is the use of the prime symbol ($'$) to refer to mutant residues, doing so, G204 and R204$'$ indicate the glycine in the wild-type and the arginine in the mutant respectively. The mutants studied are not discussed individually and extensively like the wild-type was, but are rather grouped and discussed depending upon of their structural effects. Each section will focus on a different region of G6PD (the C-terminus, the N-terminus and the enzyme core) and will present the different changes that were detected with the dynamics at different conditions.

3.6 Mutants affecting the C-terminus

G306R, G306S and A338E are mutants that were predicted to be damaging by SAAP-pred, and had an ability of inducing an increase in instability of the protein C-terminus. The instability is not dramatic to the extent of causing a global unfold of the region (or the entire protein), nevertheless there are clear indications that the local effects caused by the substitutions are influencing the C-terminal behaviour.

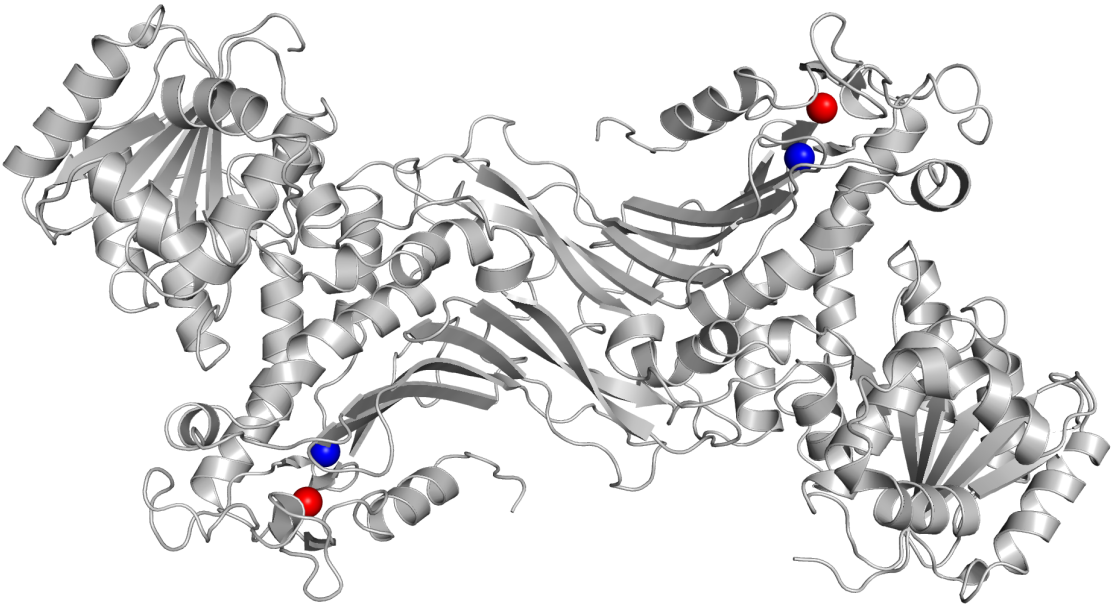
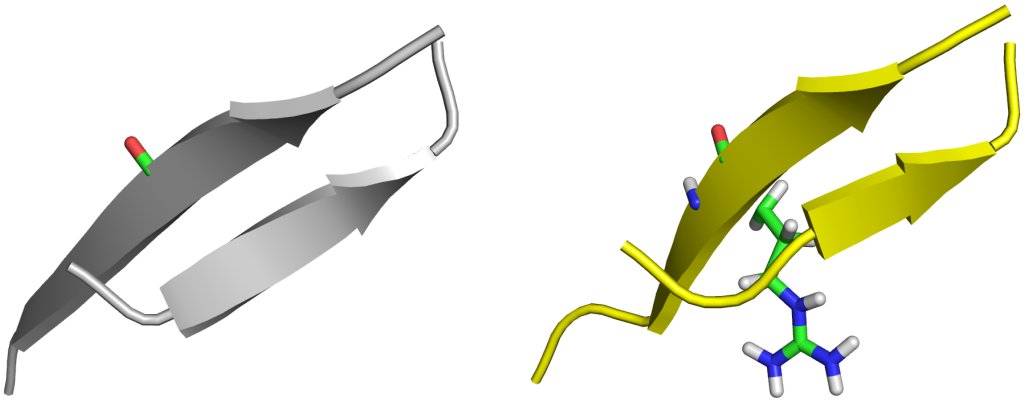


FIGURE 3.25: Location of the residues that mutated are capable of affecting the C-terminal stability. Residue 306 (G306R and G306S) is in red and residue 338 (A338E) in blue.

In G306R, a small non polar amino acid, glycine, at position 306 is replaced by a large positively charged one, arginine. The substitution was indicated by SAAPpred as damaging and the confidence of the result (0.8) was one of the highest among all the mutations considered. The residue is located in the βH strand of the central $\alpha+\beta$ 2-layer sandwich domain (Figure 3.25 in red) and when the large and charged arginine is inserted by mutmodel using the MPP protocol, it causes clashes with energy in the order of 1863.30 kcal/mol. Following the SAAPdap data, the mutation should result in an introduction of a charged hydrophilic residue in a conserved position of the core of the protein. There is no evidence of individuals with this specific mutation, but the substitution of a serine in the same position (G306S) leads to the "class II" variant originated in China and called "Seoul" [130]. Contrary to G306R, G306S introduces a mild clash (only 38.96 kcal/mol) only detected by SAAPdap in chain B of the PDB file used in the simulations. The damaging effects of the serine substitution should come

from the replacement of a mildly hydrophobic and conserved residue with a hydrophilic one. Even though the confidence for G306S of being damaging was much lower than for G306R (0.45 *vs* 0.8), it represented an good example of a mutation with a real effect on G6PD activity.

The structural effects of both mutations were similar, but more evident and enhanced in G306R. From a visual inspection of the trajectories it was not possible to detect any significant change in fold compared with the wild-type. In G306R, the arginine was oriented in a way that the hydrophilic side chain faced the solvent and because of the restricted space in the β sheet, it clashed with the isoleucine (I480) in the parallel strand (βO). The main effect of this interaction was the shortening of the βO strand itself (Figure 3.26b). In G306S, the serine side chain cannot connect the two strands, nevertheless the βH strand, where the mutation sits, shrank to half the initial length, eventually coming back to its normal length. The same happened to the opposite strand (βO), but here the strand was entirely converted into coil before refolding again. When the temperature was increased (400 K), the C-terminus of G306R was unfolding more quickly than the wild-type. After 1μs, the βO strand of the wild-type containing the isoleucine (I480) was shortened compared with the initial structure, but was still present. In the mutant, that same region was, after 500 ns, already a coil. Similarly, βH was shorter in the mutant while in the wild-type its length was maintained during the entire simulation.

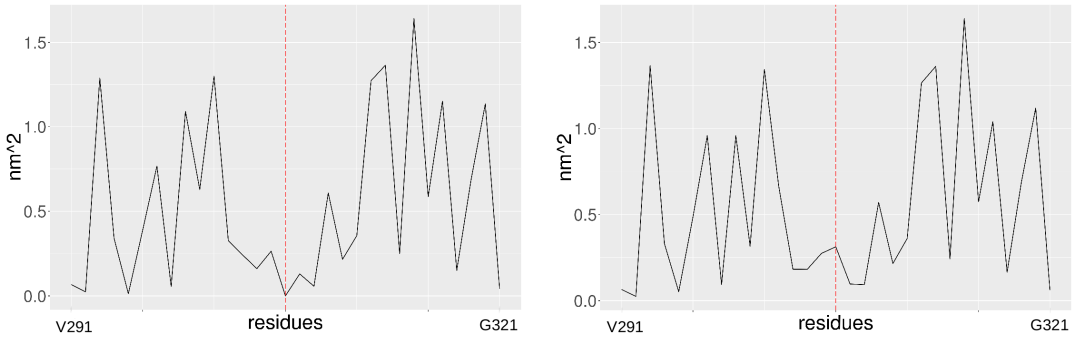


(a) (b)

FIGURE 3.26: Close view of the effect of R306 to the βO and βH strands. In (a) the wild-type, there is no contact between the two strand, while (b) in G306R, the arginine overlap I480 causing βO to shrink significantly.

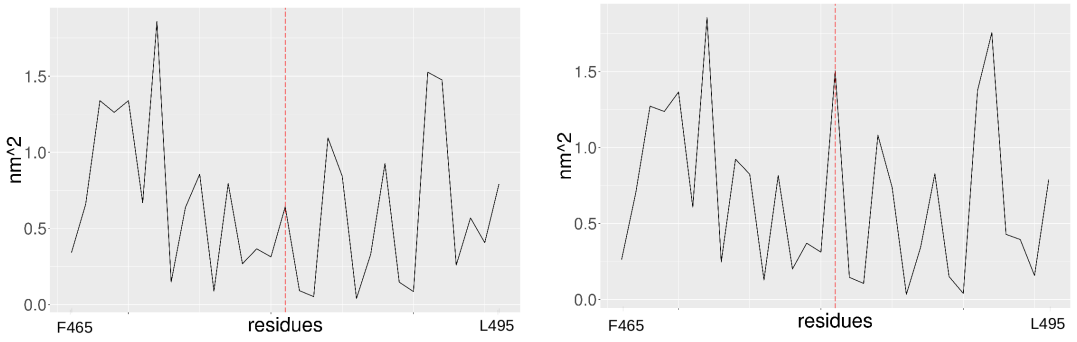
The arginine in G306R seemed to have the capability of changing the exposure to the solvent of some residues in the area. Unlike the glycine of the wild-type, R306′ can now mildly interact with the solvent (Figure 3.27b), causing R487 to double its exposure

(from 0.64 to 1.37 nm^2) and I480 to jump from 0.53 to almost 1 nm^2 (Figure 3.28b). The changes in SAS values is particularly marked in the dynamics at 400 K, where residues close to the mutation site (P312, A313 and E315) are two to three times less exposed to the solvent than the wild-type. These new rearrangements and interactions cause a 4% increase in the total potential energy compare with the wild-type (Table C.4).



(a) (b)

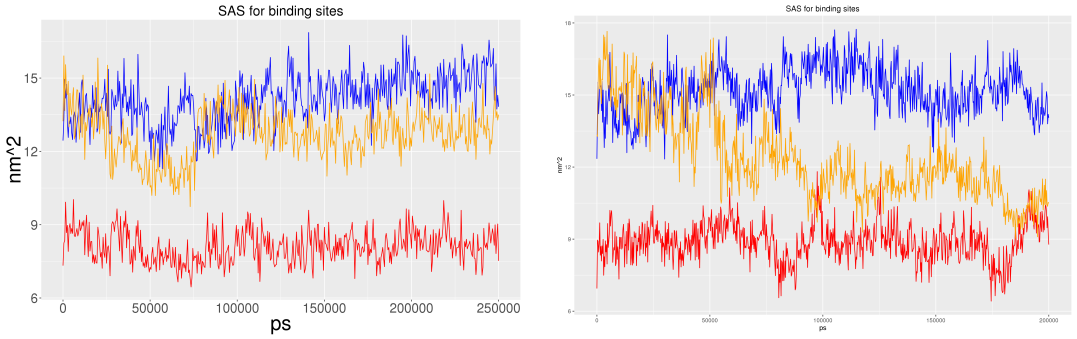
FIGURE 3.27: Average solvent accessibility area for the residues in the neighbourhood of residue 306 (dotted red line), for (a) the wild-type and (b) G306R dynamics. The different patterns in the two profiles indicates that the presence of R306′ is changing how the solvent interact with the residues in the area.



(a) (b)

FIGURE 3.28: Average solvent accessibility area for the residues in the neighbourhood of residue 480 (dotted red line), for (a) the wild-type and (b) G306R dynamics. The different patterns in the two profiles indicates that the presence of R306′ is changing how the solvent interact with the residues in the area.

In both mutants the structural NADPH$^+$ binding site had its accessible area reduced during the simulation (Figure 3.29b in orange). This appears to be the result of the higher mobility of the final part of the C-terminus that, instead of facing the outside, was found lying inside the binding site, reducing its size. Apart from the mutant site discussed above, another observable change involved the helix between residues G247 and



(a) (b)

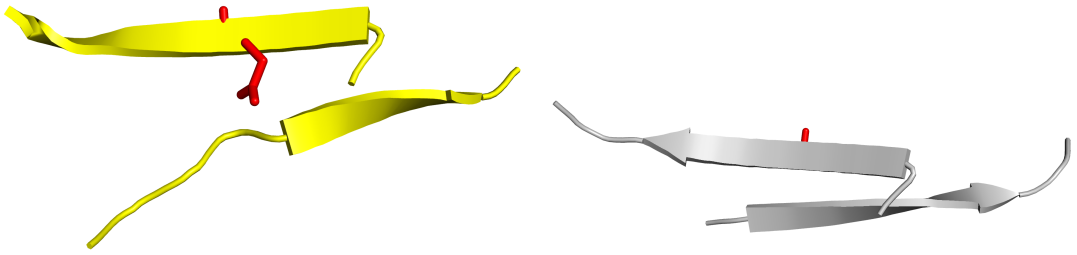
FIGURE 3.29: SAS areas of G6PD binding sites of (a) the wild-type and (b) G306S during the dynamics. The G6P site is in red, with the co-enzyme in blue and the structural site in orange.

F253 (αi) that unfolds by the end of the simulations. Even though the unfolding seems to happen at a slower rate in the wild-type, this region is not close to the mutation site and it was seen unfolding and refolding in other simulations. This suggests that a direct involvement of the mutations is unlikely, but it is possible to assume that the unfolding of the C-terminal region could have increased its stability. The rest of the residues did not present any other different or unusual behaviours compared with the wild-type. It is interesting to notice how the region from N311 to R330, that was recorded as a region of high fluctuation in the wild-type, did not move more than the wild-type. The high fluctuation of the region may have masked any other effects caused by R306′.

The dynamics of both G306R and G306S indicates that the mutations are not capable of completely unfolding G6PD in physiological conditions. Nevertheless an increase in instability in the region surrounding the mutations was observed. In R306′ , the arginine side chain does not have enough space to relax and it forces the surrounding residues (e.g. I480) to acquire different conformations, inducing a premature unfolding of the surrounding β strands. Because the C-terminus of G6PD is involved in interactions with the structural $NADPH^+$, the increase in instability caused by R306′ could weaken the affinity with the structural $NADPH^+$, eventually affecting the dimer stability. Unlike R306′-I480 in G306R, S306′ does not interact with any residues in the nearby strand resulting in a C-terminus that is less affected by the mutation destabilising effect than G306R. Also in G306S, βO shortened during the dynamics, but contrary to G306R, it eventually refolded. This may be an indication of the fact that G306S is damaging, but the damage is not strong enough to maintain the strand unfolding. The fact that G306S is a class II variant and its position close to the structural $NADPH^+$ binding site, may indicate that the disturbance caused by G306S may not be enough to destroy

G6PD structure, but is enough to trigger a series of interactions that eventually lead to a class II variant. When a more damaging mutation (G306R) is involved, similar, but heightened effects are observed.

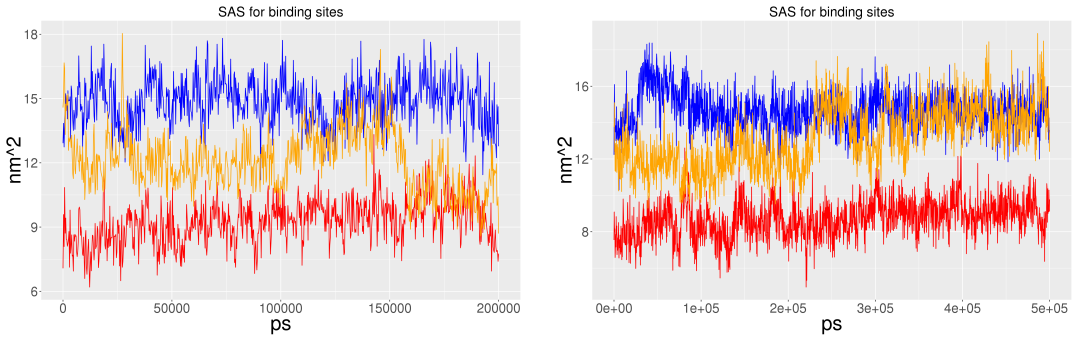
Analogously to G306R the glutamate in A338E (Figure 3.25 blue) causes the βH and βI strands to separate until the N-terminal segment of βH (T336 and F337) eventually unfolds (Figure 3.30b). In the wild-type, A338 is deep in the core of the enzyme, shielded from any contacts with the solvent. Instead, in A338E, the rearrangements of the area around the mutation expose E338′ to the surface.



(a) (b)

FIGURE 3.30: (a) In A338E, the initial part of βH unfolds under the effects of the glutamate (red) in the adjacent strand. (b) The wild-type is presented as a reference.

The size of the structural $NADP^+$ (Figure 3.31b) noticeably increases compared with the wild-type (Figure 3.31a). This difference is only 1 nm^2 if calculated as the average size value over the trajectory, but contrary to the wild-type, in which the size decreases to values of 10 nm^2 at the end of the simulation, in A338E the same area grew to values oscillating around 15 nm^2 (Table C.4). It is probable that E338′ is involved in this behaviour.



(a) (b)

FIGURE 3.31: SAS areas of G6PD binding sites of (a) the wild-type and (b) A338E taken over the entire trajectories. The G6P site is in red, with the co-enzyme in blue and the structural site in orange.

3.7 Mutants affecting the N-terminus

Similarly to what happened in the C-terminus, when a damaging mutation occurs in the N-terminus of G6PD, very local effects connected to region instability are generated. Among the mutants studied, Y70H, R136C and A^- are representative examples of the damage caused. The distortions to the N-terminus can be grouped into mutants affecting the top helices of the Rossmann-like domain and mutants destabilising the co-enzyme binding site. The A^- mutant (Figure 3.32 in blue) belongs to the first group, while Y70H (Figure 3.32 in red) and R136C (Figure 3.32 in yellow) to the second.

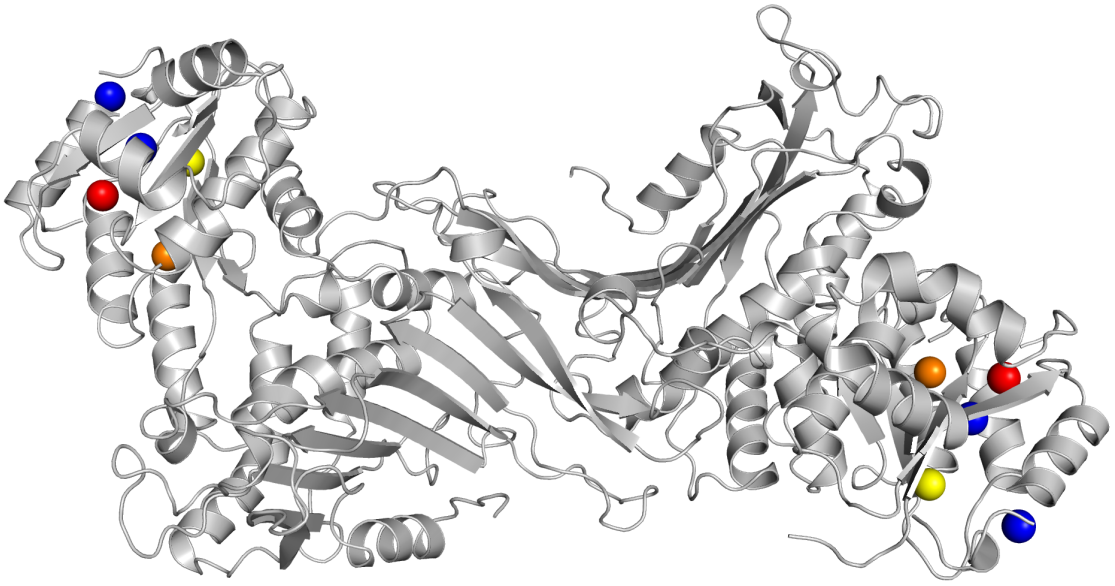
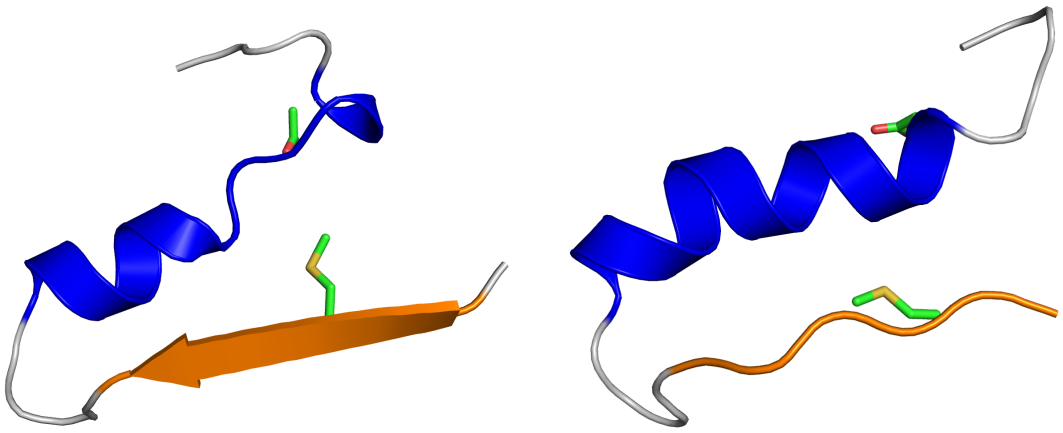


FIGURE 3.32: Location of the residues that mutated are capable of affecting the N-terminal stability. A^- is in blue, Y70H in red, L140P in orange and R136C is in yellow.

The A^- mutation is one of the few multiple missense G6PD variants. The combination of V68M and N126D results in a class II mutant [131]. This G6PD variant probably originated from the non-deficient single-mutant A variant, which only has the substitution of the valine (V68M). Even though SAAPdap found no significant local structural effects, and only one of the mutations (V68M) that constitute this variant was individually predicted as damaging by SAAPpred with a very low confidence (0.02), A^- is considered of great importance because it is one of the most common variants in African, or African ancestry, populations [132]. Previous studies suggested that the depressed activity of A^- is probably caused by interactions between the two mutated residues, leading to an increase in hydrophobic area exposed to the solvent [133]. The mutant does not have different affinity for G6P, but a lower affinity for NADP was observed in A^- , compared with the wild-type [132].

The dynamics show that D126′ destabilised αc, which unfolds when M68′ is oriented in a way that its side chain points to αc. The βB strand, which is where M68′ is located, is stable when αc unfolds and unstable when αc is folded (Figure 3.33). These conformational changes indicates that, unlike the other mutants studied, in the case of A^- it may be possible to detect a clear structural effect of the mutation on the Rossmann-like domain of the enzyme. Even though the mechanism of action is not very clear, this study was also capable of confirming the propensity of A^- to favour β-sheet structures as proposed in a previous study [133]. When the M68′ side chain points to D126′ (Figure 3.33a), helix αc unfolds and there is a clear change in exposure to the solvent in the area. When the M68′ side chain sits parallel to the β strand of which it is part (βB), αc maintains its conformation and it is βB which unfolds (Figure 3.33b). In both cases, M68′ was always in the core of the domain, protected from any interactions with the solvent. This reciprocal movement involved only the residues close to the mutations, as the rest of the region did not considerably diverge from the wild-type.

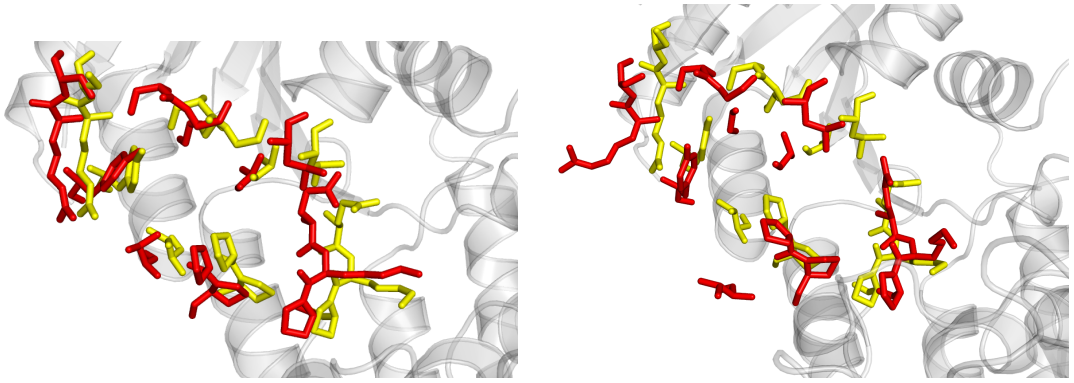


(a) (b)

FIGURE 3.33: Relationship between αc (blue) and βB (orange) in A^- dynamics. (a) When the M68′ side chain points to the αc helix, αc unfolds, while (b) when the M68′ side chain sits parallel to the βB strand, αc refolds and βB unfolds.

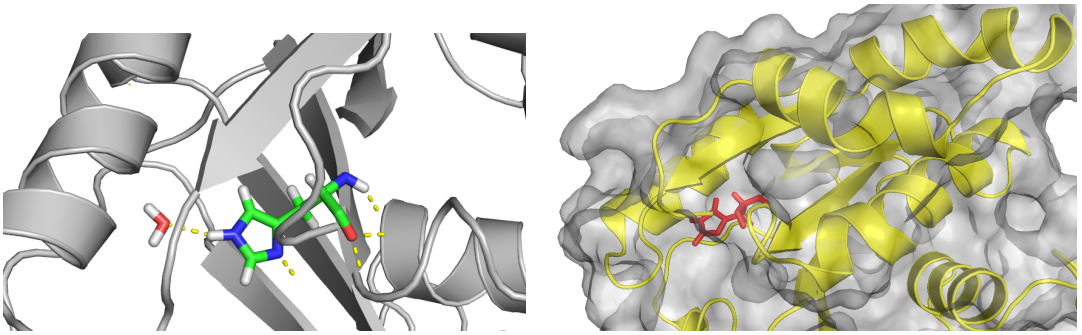
Contrary to A^- , the mutations Y70H and R136C distort the co-enzyme binding site. Mutant Y70H occurs in the βB strand, in between two other mutations studied: L140P and A^- . In terms of steric effects, a histidine is not dramatically different from a tyrosine, but because of the charged side chain, SAAPdap indicates the burial of a charged hydrophilic residue in the core of the protein and the disruption of hydrogen bonds as possible local structural effects. Y70H, which was predicted by SAAPpred as being damaging with a confidence of 0.67, is a class II G6PD variant called Namouru, which was first reported in the small group of Pacific islands of Vanuatu [134]. Also R136C is a class II variant (Valladolid), and it is characterised by normal heat and

kinetic properties [135]. Zara *et al.* [135] suggest that the low RBCs residual activity may be the result of the reduced affinity for the enzyme substrate. The native residue, R136, sits in a cleft inside the Rossmann-like domain and forms hydrogen bonds with the surrounding residues, in particular with the backbone of G131. SAAPdap indicates the disruptions of hydrogen bonds and the removal of the charge as possible local effects. The replacement with the cysteine breaks these bonds, causing further instability in the region. R136C was predicted to be damaging by SAAPpred with a confidence score of 0.45. Contrary to the arginine, C136′ cannot be involved in any hydrogen bonds and all the arginine bonds with any αc residues (backbone of G131 and N122) or the side chain of C158, located in the loop between αd and βE, were lost. The mutation of C158Y is also a class II variant, called Shenzen. Because C158Y was found capable of affecting the co-enzyme binding stability, it is possible that interactions of C136′ with C158 could produce a similar effect. The two residues are too far from each other (6Å measured with PyMOL) to create a disulphide bridge. Both mutants cause the partial unravelling of helix αe (L177-L190) but with different mechanism and effects. In the wild-type, the arginine (R136) side chain is involved in hydrogen bonds that stabilise the adjacent αc helix. In R136C, these bonds are not maintained and helix αe is free to enter and distort the co-enzyme binding site when it unfolds. This happened also in the wild-type after almost 300 ns, but in R136C, the unravelling happened after only 100ns. The helix eventually refolded, but it was enough to have an effect on both size and geometry of the co-enzyme binding site. The fact that the unravelling is quicker than all three wild-type replicas indicates that this could be a significant effect of C136′. When helix αe unfolded, the structured parts of the helix moved toward the centre of the co-enzyme binding site. The *cis* conformation of Pro172 restricted the movement and forced these residues onto the outside, resulting in a more compact and disordered binding site (Figure 3.34b). In accordance with what was observed by Zara *et al.*, the distortion may be big enough to change the co-enzyme affinity in the mutant. In Y70H the resulting co-enzyme is bigger than the wild-type. The mutation increases the motility of the exposed helices constituting the N-terminal Rossmann-like domain eventually allowing some water molecules to enter the mutant core (Figure 3.35a). Residue H70′ had a small increase in area exposed to the solvent (calculated as the mean over the trajectory), suggesting that H70′ was exposed in some frames of the trajectory, but for most of them, it kept its position buried in the protein core (Figure 3.35b). The movements of the N-terminus destabilises the helix αd even more, such that it eventually unfolds, speeding up the distortion of the co-enzyme binding site. However, the exact mechanism by which H70′ causes this chain of events was difficult to detect.



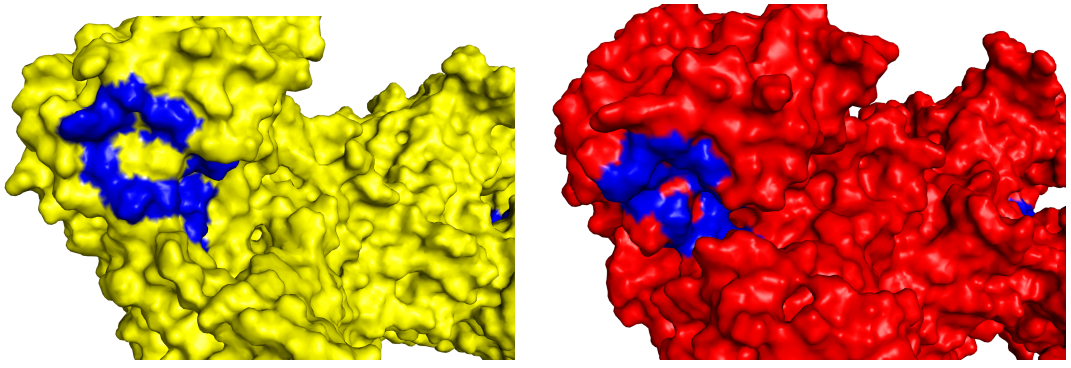
(a) (b)

FIGURE 3.34: Change in the geometry of the co-enzyme binding site in two replicas of R136C. The residues that binds the NADP in the wild-type are coloured in yellow, while the same residues are coloured in red to outline the different positions that is found in R136C.



(a) (b)

FIGURE 3.35: (a) H70′ interacting with a molecule of water during the dynamics. (b) The location of H70′ (red) buried inside the 'NAD(P)-binding Rossmann-like' domain of Y70H.



(a) (b)

FIGURE 3.36: Deformation of the co-enzyme binding site in Y70H (a) compared with (b) the wild-type.

3.8 Mutants affecting the core

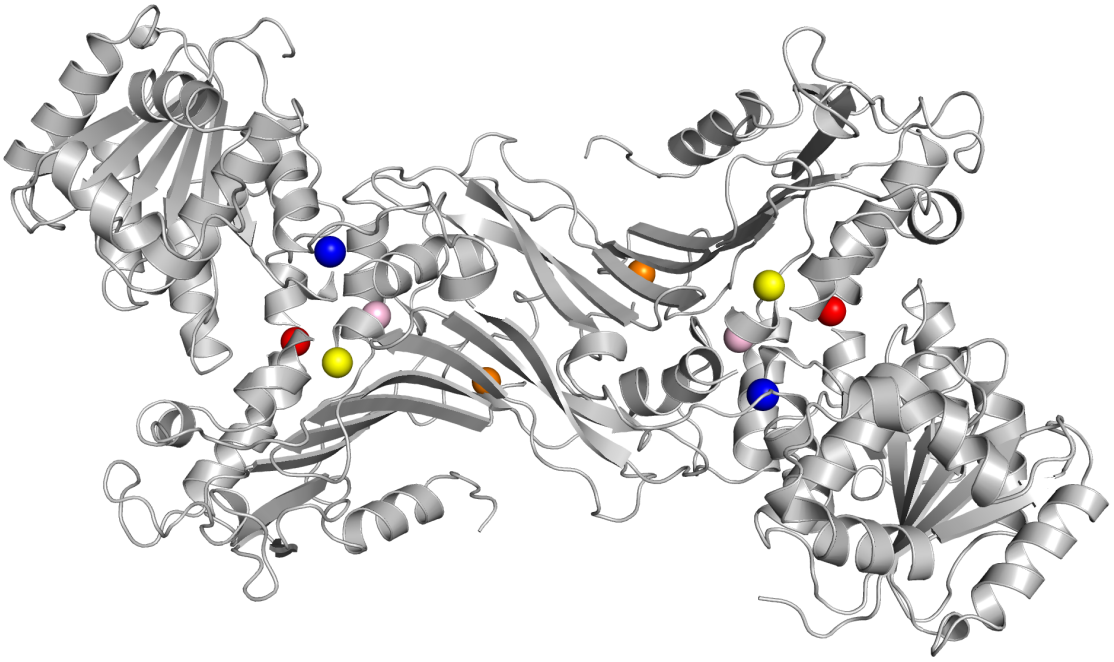
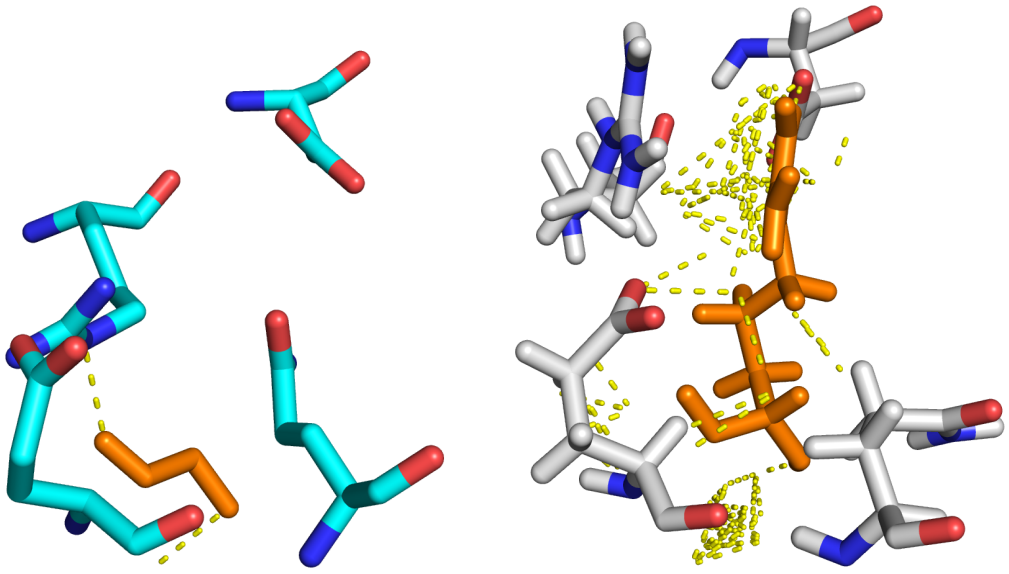


FIGURE 3.37: Location of the residues that mutated are capable of affecting the core stability. G204R is in blue, A461T in red, R227Q in orange, L264R in yellow and C269Y is in pink.

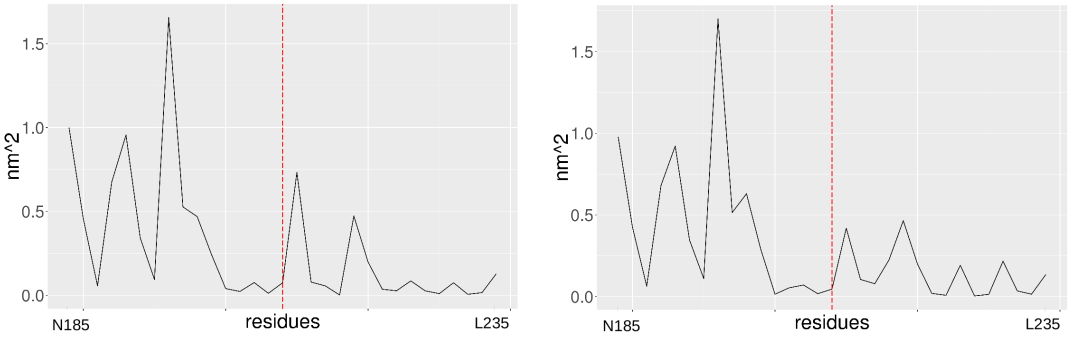
Similar to G306R, the SAAPdap analysis suggests that the presence of R204′ introduces, in G204R (Figure 3.37 blue), a charged hydrophilic residue in the core of the protein, with clashes in the order of 419.62 kcal/mol. The mutation site is in a small helix, spanning four residues, three of which (His 201, Tyr 202 and Lys 205) are involved in hydrogen bonding with the substrate. As a consequence the mutation, not only damages the structure by introducing a large charged residue, but it also directly affects the binding with the substrate. SAAPpred predicted G204R to be damaging with a confidence of 0.8. In the crystal structure of the wild type (PDB code 2bhl), G6P was bound to the protein and in both chains the region H201-H205 was folded into a helix. During the simulation, it was observed that for both the wild-type and G204R, the small helices tended to alternate between folded and unfolded states. Because the simulations were performed in the absence of G6P bound to the chains, it may be that G6P binding stabilises the helix, and that its absence may increase the disorder of the area, allowing the short helix to unfold/fold. The arginine side chain is much larger than the glycine one, but the arginine side chain is capable of positioning in a cleft above the residue. In this position, during the simulation, the nitrogens of its side chain are capable of interacting with the residues of the surrounding helices: R439, D443, E206 and Q209 (Figure 3.38b).



(a) (b)

FIGURE 3.38: In yellow are represented all the (a) polar contacts in which G204 is involved in the wild-type, and (b) that of the R204′ side-chain in G204R. R204′ interacts with the surrounding residues (R439, D443, E206 and Q209) reducing the motility of the region.

In this position, R204′ can now interact with the solvent, forcing the rearrangement of the surrounding residues with D200, which was forced to a different position, resulting in a decrease in solvent accessible area (Figure 3.39b). A similar decrease happened for the residues interacting with the arginine side chain: R439, D443, E206 and Q209.



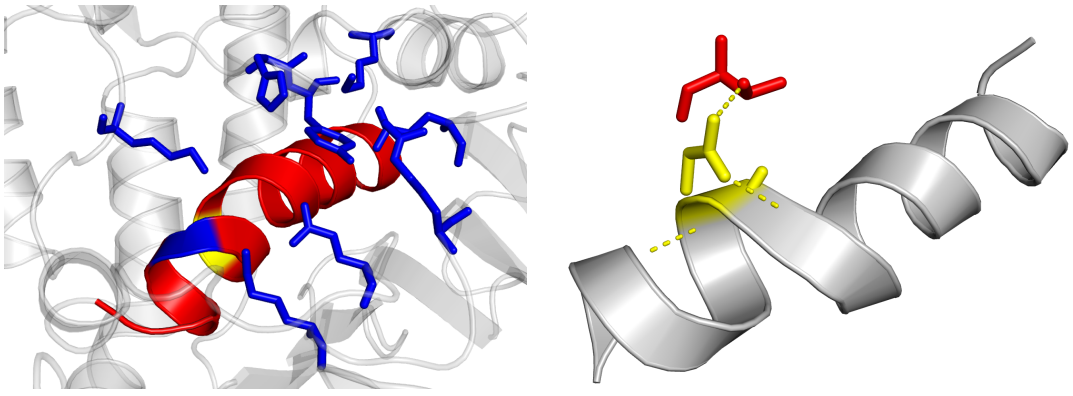
(a) (b)

FIGURE 3.39: Average solvent accessibility area for the residues in the neighbourhood of residue D200 (dotted red line) for the (a) the wild-type and (b) G204R dynamics. The different patterns in the two profiles indicates that the presence of R204′ is changing how the solvent interacts with the residues in the area.

These interactions were always present during the simulations, suggesting that they

might act as constraints to the arginine movement, resulting in an helix with less capability of folding/unfolding. In Bautista *et al* [82] it was observed that, in the G6PD variants involving K205 (K205T and K205R), the mutation causes a severe drop in k_{kat}, without affecting the K_m, suggesting that K205 is dispensable for substrate binding, but essential for catalysis. G204R may act in a similar manner: the stiffened helix does not cause any major structural disruptions, but it changes and maintains a different geometry inside the binding site.

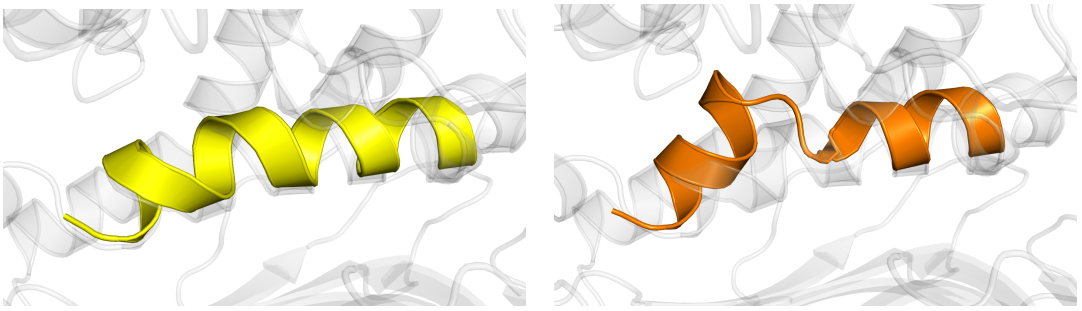
Another damaging mechanism affecting the core is represented by the distortion of the αi helix which form the supporting base of the G6P binding site. The mutant A461T is a G6PD variant which was first isolated in individuals from the Yunnan chinese province. Xiaoquin *et al.* [136] limited themselves to the detection of the mutation, without providing any enzymatic or structural information. In the pool of people that was studied, the authors found a mixture of other variants common in China, allowing them to suggest similarities with both the Canton and Kaiping variants (both class II). In the reference database of G6PD variants [101], A461T was classified as "NR" (non reported). Residue 461 is contained in the central section of the long αn helix on the back of G6PD (Figure 3.37 in red) and, contrary to the alanine, the threonine (T461′) can interact with the glutamine (Q261) in the adjacent αi helix (Figure 3.40a). Helix αi spans 17 residues from I255 to A272 and constitutes the supporting base of the G6P binding site (Figure 3.40b). During binding, helix αi is ordered and D258 and H263 interact with G6P,while in absence of it, the last section of αi is less ordered and disruptions in the helix was observed several times in other simulations. The threonine could interfere with this mechanism, affecting the way the binding site overall behaves.



(a) (b)

FIGURE 3.40: (a) Bulge on the αi helix caused by Q261 (yellow) interacting with T461′ (red). The dotted lines are the polar contacts of T461′ as detected by PyMOL. (b) Close up view of helix αi (red). The residues binding G6P are coloured in blue, while Q261 is in yellow.

The distortion of the base is increased in L264R, which was capable of dramatically reducing the size of the G6P binding site. This was possible because L264 is a central residue in helix αi that constitutes the base of the G6P binding site. When the leucine was replaced by the arginine, αi broke into two independent pieces (Figure 3.41b). The first ten residues (I255-H263) of helix αi deformed quite often in previous simulations, but L264R was the first mutant in which there is a clear separation of the two segments of the helix.

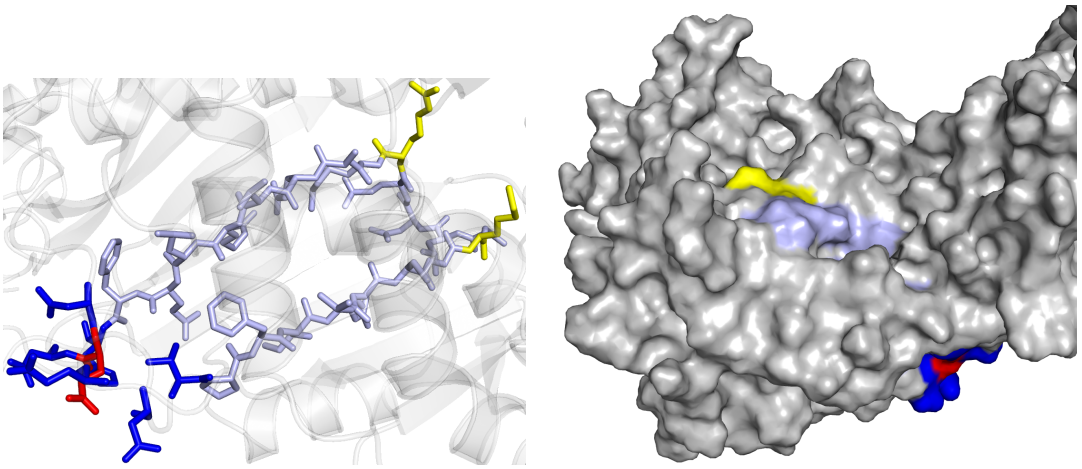


(a) (b)

FIGURE 3.41: Representation of the αi helix as found in L264R (a) at the beginning and (b) at the end of the dynamics. The wild-type structure is superimposed in grey.

A particular case is represented by the class II variant known as 'Mexico City' (R227Q) [137], in which the effects of Q227′ can travel from the surface of the protein to its core, directly affecting G6P binding. The mutation is in the loop (P223-A231) that connects together the βG strand of the central β region with the αg helix that precedes the triad of residues that binds G6P (H201, H202 and K205). Q227′ forces a change in position of the residues forming a patch around the mutation. The surface residues are connected by a sequence of residues that reaches the core of the protein before going back to the mutation site (Figure 3.42a). The movements induced by Q227′ are capable of travelling from the surface to the core and misalign two important binding residues: K360 and R365.

During the dynamics it was observed that Q227′ is surrounded by six residues (N226, D228, E347, D350, V352 and H374) forming a patch of residues similarly exposed to the solvent. Among these residues, all but D350, which saw its exposed area growing by half, did not change significantly the way they interacted with the solvent. Even though these changes may not in themselves explain the damaging effects of R227Q, they describe very well a mechanism that could explain how the damage is propagated throughout the G6PD structure. The difference in side chain between the arginine and the glutamine forces the surrounding residues to redistribute around Q227′. On the surface, the residues forming a part of the patch surrounding Q227′ (E347, D350 and V352)



(a) (b)

FIGURE 3.42: (a) Representation of the residues surrounding Q227′ (red). In yellow are K360 and R365, which are part of the G6P binding site, and in blue all the residues surrounding the mutation (N226, D228, E347, D350, V352 and H374). The sequence of 20 residues that connects the G6P binding site to the surface of G6PD is in light blue. (b) Surface view of R227Q. Q227′ is in red, with the residues surrounding the mutation (N226, D228, E347, D350, V352 and H374) in blue. K360 and R365, the residues binding G6P are in yellow, while in light blue are indicated the residues that connect the surface patch and the G6P binding site.

are connected to H374, which completes the surface patch, by a sequence of 20 residue (Figure 3.42). Among the 20 residues, two are known to bind G6P in its binding site: K360 and R365 (Figure 3.42). The movements caused by the mutation can propagate, through the sequence of 20 residue, to the binding site, orienting K360 and R365 side chains in a non-optimal position for G6P binding. Evidence of this mechanism is that during the dynamics of R227Q, R365 is less exposed to the solvent than in the wild-type. If what is proposed is true, it could explain why R227Q is a class III variant and R227L (a larger difference from arginine) is a class II variant. In R227L, the hydrophobic leucine is forced away from the surface causing larger structural rearrangements in the area.

3.9 Mutants with additional behaviours

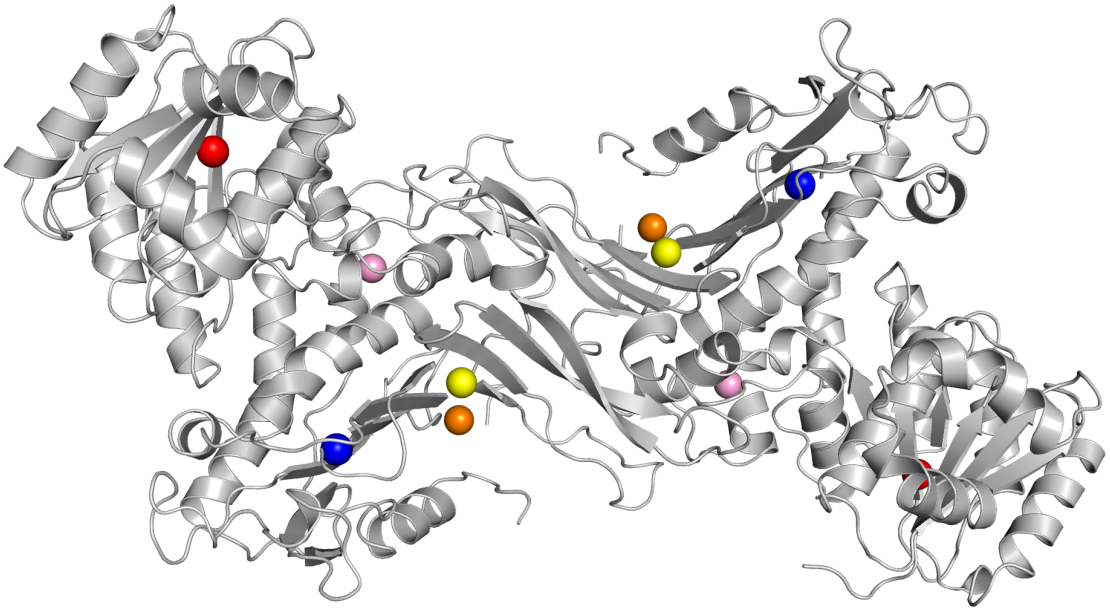
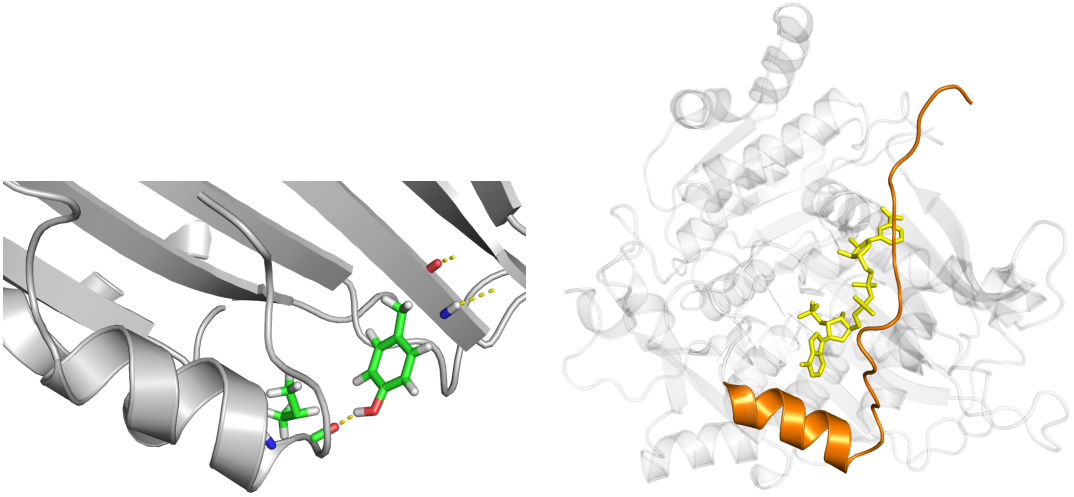


FIGURE 3.43: Location of the residues that differ from the mutants described in the previous sections. G359R is in blue, L137P in red, E287K in orange, R370W is in yellow and C232Y is in pink.

Some mutants were initially selected for the study, but it was not possible to collect more than a single trajectory and they were discontinued. The reasons vary from mutants that did not present differences from the wild-type dynamics (L137P or E287K), mutants for which it was not possible to explore the damaging mechanism further (C232Y, see below) and finally mutations that were selected for study towards the end of the project when time was limited (G359R and R370W). This section contains a brief description of their trajectories and their possible mechanisms of action.

Mutant C232Y has a tyrosine replacing a cysteine at the beginning of the βG strand. It causes a severe decrease in G6PD activity (class I), thermal instability, low affinity for NADPH and product inhibition by NADPH itself [138]. SAAPdap indicates that the tyrosine side chain causes clashes ranging from 126.12 to 2130.24 kcal/mol, and C232Y was predicted as damaging by SAAPpred with a confidence of 0.49. During the single 150 ns trajectory, it was noted that the tyrosine of C232Y was interacting with V499 (Figure 3.44a). V499 is both the end of the αl helix and the beginning of a long loop that reaches the end of the G6PD sequence. This long loop allows $NADP^+$ to enter the structural $NADP^+$ binding site and stabilises its stay by providing three binding residues (Figure 3.44b). It is possible that, in C232Y, the tyrosine could cause the low affinity by reducing the mobility of the long loop in the C-terminus, resulting in an inaccessible

NADPH structural site. Even though C232Y presented an interesting mechanism, the truncated structure (see Section 3.3) did not allow to study the changes in stability of the C-terminal loop, and the study of this mutant was halted.

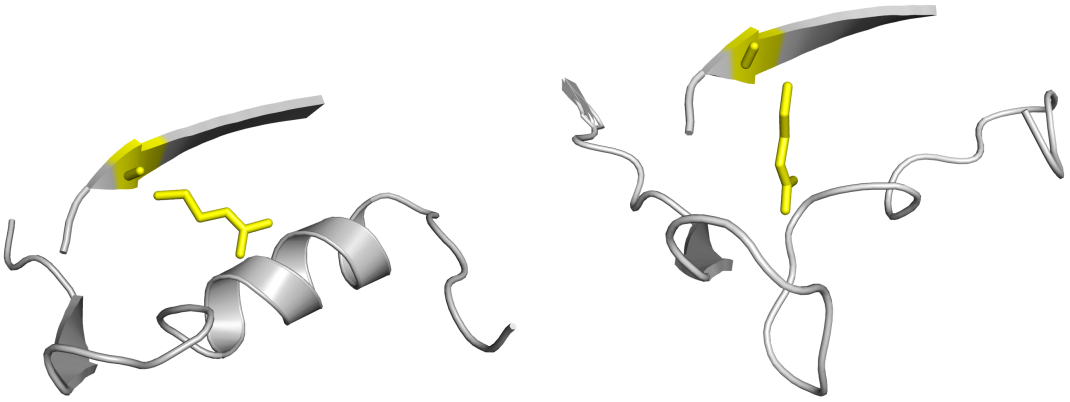


(a) (b)

FIGURE 3.44: (a) The Tyrosine interacting with V499 in the adjacent helix. (b) The NADPH binds to the the non truncated wild-type (2bh9.pdb). The long loop at the end of G6PD is coloured in orange, while Y232' is in red.

G359R is another unknown variant in which a glycine is replaced by an arginine. Similar to both G204R and G306R, G359R has a high confidence of being damaging (0.79 as predicted by SAAPpred) and, in the 500 ns long trajectory at 400 K, it was possible clearly to observe the damage caused by this mutation. The arginine side chain did not have enough space to relax itself, causing all the surrounding structures to grow apart and relocate. Eventually, the αo helix and the surrounding helices unfolded (Figure 3.45 in yellow), leaving only the β strands intact. Furthermore, the structural $NADP^+$ binding site was found to be much bigger than the wild-type. In the wild-type, it ranged between 12 and 15 nm^2, while in G359R, it was constantly above 15 nm^2 with peaks of 20 nm^2. Even though the arginine side chain is oriented away from the G6P binding site, the effects of all the rearrangements caused the G6P binding site to shrink. The shrinkage did not have the same magnitude as the expansion of the structural $NADP^+$ binding site, nevertheless it was visible.

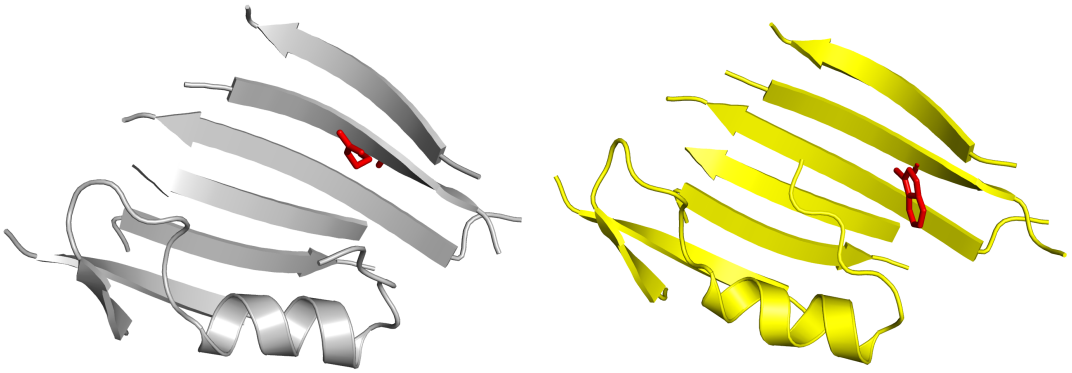
R370W is a G6PD variant not yet described in humans, and it occurs in a residue that interact with the structural $NADP^+$ to stabilise G6PD. In the wild-type, the R370 side chain stays parallel to the βK strand, leaving the binding site free (Figure 3.46a). In R370W, the tryptophan assumed a position in the middle of the binding site. This



(a) (b)

FIGURE 3.45: Unfolding of the αo helix in G359R. (a) At the beginning and at the (b) end of the dynamics. R359′ is in yellow.

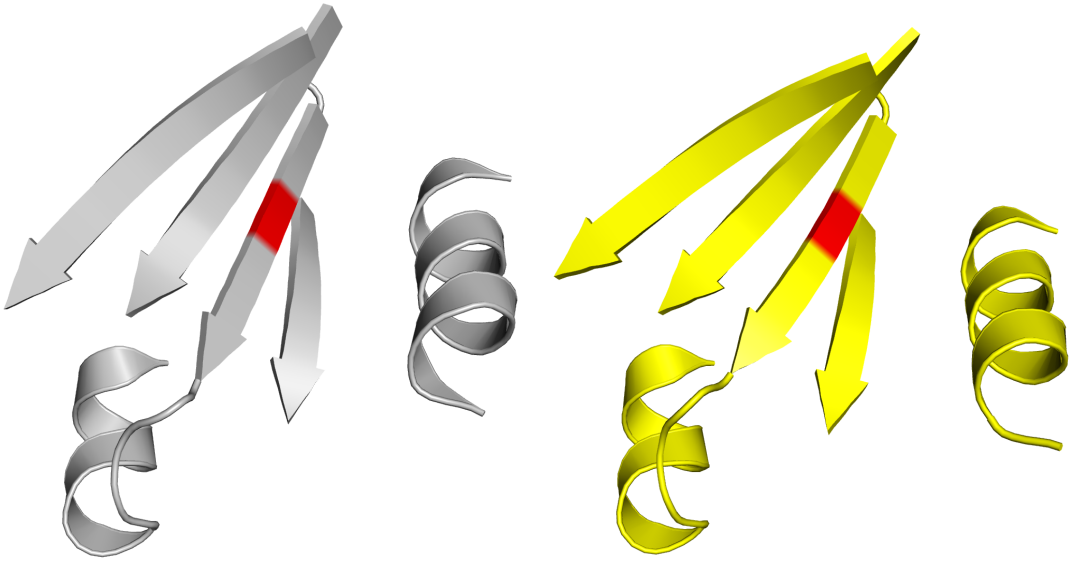
movement made it close to Y503 on the opposite loop, causing the structural $NADP^+$ binding site to reduce its size (Figure 3.46b).



(a) (b)

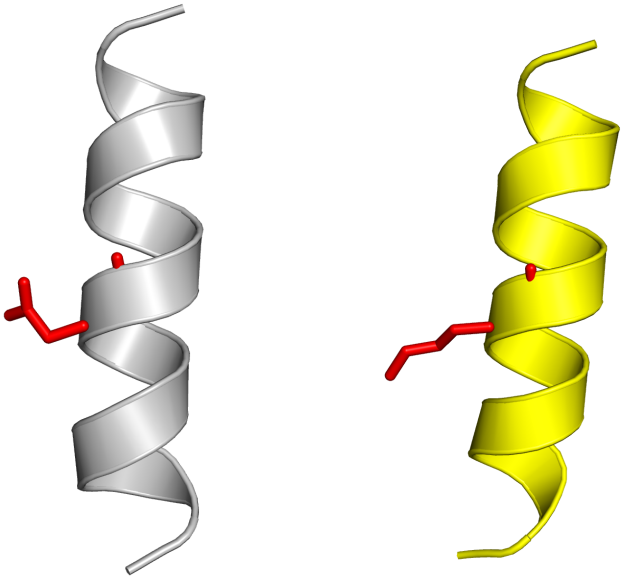
FIGURE 3.46: Representation of the area around residue 370, in red, as seen in (a) the wt and in (b) R370W.

L137P and E287K are both mutants that did not show any sign of a different behaviour compared to the wild-type (Figure 3.47 and Figure 3.48). L137P was predicted to be damaging by SAAPpred with a confidence of 0.78 and introduces clashes at a site where the backbone torsion angles could not accommodate a proline, while the introduction of a lysine in E287K disrupts hydrogen bonds of a residue involved in binding and was predicted to be damaging by SAAPpred with a lower confidence than L137P (0.69).



(a) (b)

FIGURE 3.47: Representation of the area around residue 137, in red, as seen in (a) the wt and at (b) the end of the dynamics of L137P. The proline has no effects on the surrounding area.



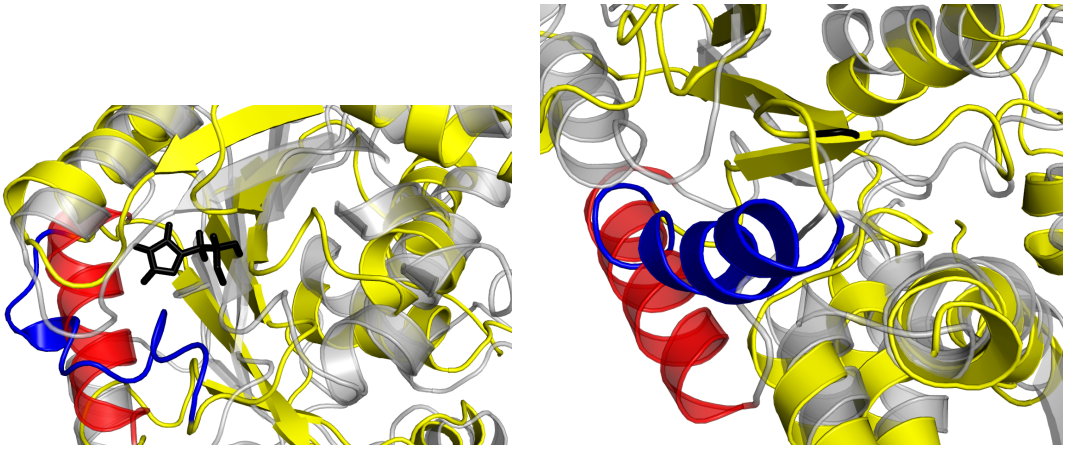
(a) (b)

FIGURE 3.48: Representation of the area around residue 287, in red, as seen in (a) the wt and at (b) the end of the dynamics of E287K. The lysine has no effects on the helix.

3.10 A note on Pro172 in the mutants

Section 3.4.6 described the characteristics and behaviour of Pro172 during the dynamics of the wild-type. In the simulations at temperatures below 500 K, Pro172 always stayed in the *cis* conformation and *trans* values of the ω angle were recorded only at 500 K. The mutants were not studied at 500 K, and there were no records of any *cis-trans* shift in their dynamics.

During the analysis of G204R and Y70H, it was noted a feature of the V169-L176 loop, which could be evidence of the structural importance of Pro172. This observation came after noticing that, during the dynamics, the αe and αd helices tend to invade the co-enzyme binding site, resulting in its deformation. In Y70H for example, the start of movement of the N-terminus coincides with the unfolding event of the αd helix (Figure 3.49a). The absence of this helix allowed the αe helix to push forward towards the co-enzyme binding site (Figure 3.49b), reducing the co-enzyme binding site area. Because the start of movement of the N-terminus coincides with the unfolding event of the αd helix (Figure 3.49b in red and blue) it is possible that the *cis* conformation of P172 forces a twist in the V169-L176 loop, which was helping to keep the preceding helix (D176-L190) away from the co-enzyme binding site. The absence of this helix allowed the αe helix to push forward towards the co-enzyme binding site (Figure 3.49b), distorting the co-enzyme binding site area (Figure 3.36). Possible cis-trans isomerisation of Pro172 is further explored using metadynamics in Chapter 4.



(a) (b)

FIGURE 3.49: The role of Pro172 (black) in maintaining αe away from the co-enzyme binding site. (a) αd in both folded (red) and unfolded (blue) conformations. (b) In the absence of αd, αe (red and blue) moves inside the co-enzyme binding site.

3.11 Discussion

TABLE 3.3: The main mutants studied, together with a brief overview of their structural effects observed during the simulations. N-term refers to the Rossmann-like domain, where the co-enzyme binds, while C-term refers to the terminal region of the central Reductase domain, where the structural $NADP^+$ binds. G6P stands for the area around the G6P binding site.

Mutant	Class	Location	Effects
G306R	-	C-term	premature unfolding of the C-term
G306S	II	C-term	destabilising effects on the C-term
G204R	-	core/G6P	restrains on the helix movements
L140P	-	N-term	no noticeable effects
Y70H	II	N-term	distorted co-enzyme binding site
C269Y	I	core/G6P	no noticeable effects
A338E	-	C-term	premature unfolding of the C-term
R227Q	III	surface	change in SAS exposure around G6P
A461T	NR	G6P	deformed helix in G6P
R136C	-	N-term	distorted co-enzyme geometry
A^-	III	N-term	unfold/refold of the areas surrounding the mutations
L264R	-	G6P	breakage of the helix at the base of G6P

FIGURE 3.50: The regions which were indicated by the dynamics as regions of high fluctuation are coloured in orange (one chain only).

All the work presented in this chapter is based on the assumption that the depressed activity registered in the G6PD variants is the result of specific structural changes in the G6PD structure, which are shared among the mutants and are therefore detectable by a

small set of antibodies. To test the truthfulness of the assumption, the likely structural effects of G6PD mutations were studied first with the SAAPtools and then with extensive MD simulations on a small subset of damaging (predicted and not) mutants. The study of the effects of these mutations has proven to be challenging, mainly because of the difficulty in collecting extensive simulation data and in the overall interpretation of the results. The dimer of G6PD, with 958 residues and almost 8000 atoms, is considered a large system to be studied with molecular dynamics. A single 500 ns trajectory can take months of simulation time and, as a direct consequence, in several cases it was not possible to study the same mutant with different independent simulations (replicas). Furthermore the absence of abundant structural information on G6PD and its variants made the choice of mutations to study and the extraction of results more difficult. The core of this project was the structural analysis of G6PD variants to locate some behaviours shared among them and not present in the wild-type. Extensive all-atom MD simulations, had indicated that the mutations are not capable of causing severe structural alteration to the protein structure. Instead, each case had its own small and local rearrangements that sometimes made the connection between structural damage and depressed activity phenotype possible (Table 3.3).

At the beginning it was thought that a high quantity of unstructured regions (coils, loops and turns) in the protein may play a key role in maintaining G6PD function, acting as "cushions" around the important structures, the coils could have been capable of absorbing the damaging effect of the mutations. The hypothesis that G6PD had a significant amount of coils and loops, compared to other proteins, was tested by comparing the percentage of unstructured regions found in G6PD over the protein structures in the Protein Data Bank. Initially all resolved structures were considered, but later the analysis was repeated only for dehydrogenase structures first and enzymes known to be vital to sustain life then. To reduce the noise and refine the results (e.g. partial structures, small peptides and accessory structure) the analysis was performed only on the structural domains recognised by CATH [139]. The representative structure ($\leq$ 35 % sequence identity) for every homologous superfamily was selected, leaving out NMR structures and proteins shorter than 30 amino acids. The secondary structure information was then extracted and counted. The percentage of coils in the N-terminal Rossmann-like domain is 19%, with 44% of alphas and 37% of betas. In the second dihydrodipicolinate reductase domain, where G6P binds, the percentage of coils drops to 11%, with 49% alphas and 39% betas.

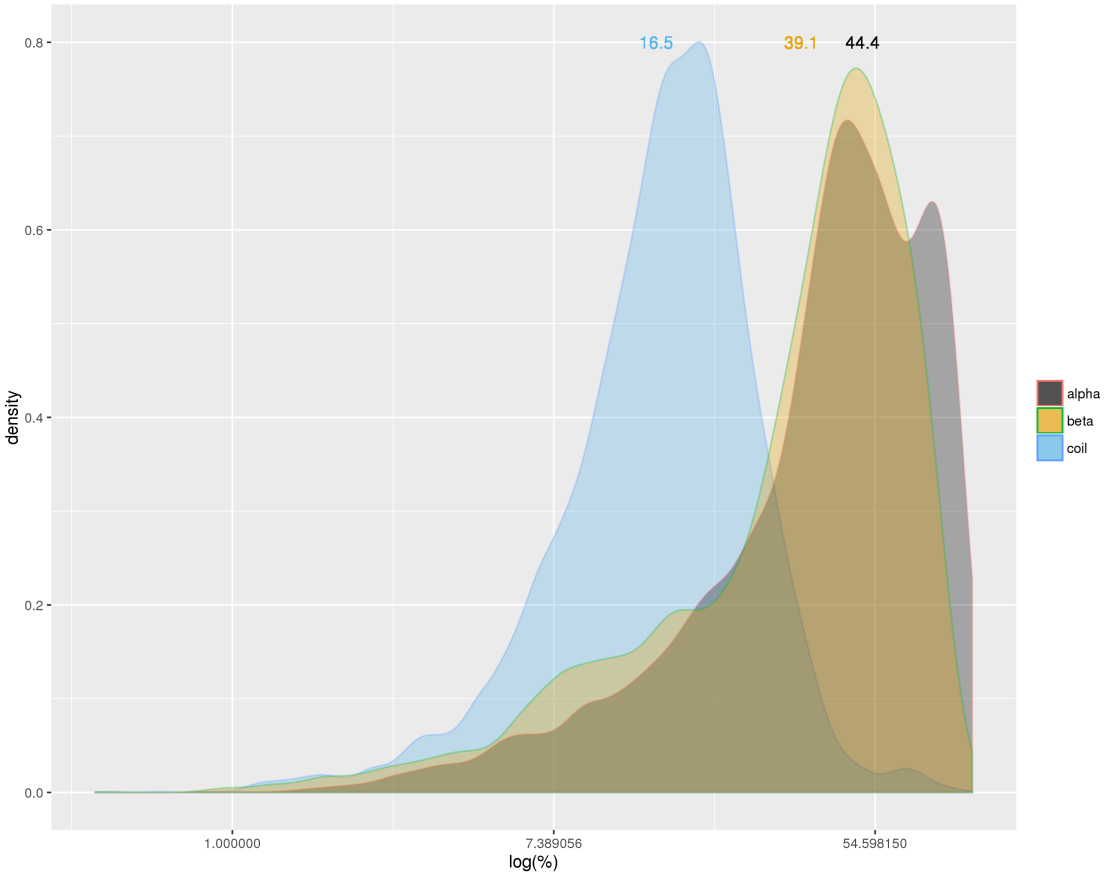


FIGURE 3.51: Distribution of the three different types of secondary structure of the CATH structural domains. The averages of each group are shown on top of each distribution. The x-axis is in logarithmic scale.

The alpha and beta distributions have long tails, so the nonparametric Mann-Whitney test, which does not require normal distributions, was used to check if the percentage of coils in G6PD was significantly less than or greater than the same value in another sample (distribution of the CATH domains (Figure 3.51). The Rossmann-like domain in G6PD has almost 3% more coils, contrary to the dihydrodipicolinate reductase domain that has 4% less. The Mann-Whitney test on the data suggests that G6PD may have a significant difference in coils in both domains (p-value $< 2.2\times10^{-16}$), but it is unlikely that a connection between the percentage of coils and protection against the damaging effects of mutations really exists. In fact, 4% only represents 8 and 11 of the residues that constitute the two G6PD domains respectively. In an attempt to show any correlation between the SAAPpred performance and the WHO class classification of G6PD variants, all the existing mutations were associated with their confidence of being damaging obtained from SAAPpred. The resulting distributions (Figure 3.52) indicate that SAAPpred is capable of correctly detecting the moderate to severe G6PD mutations (class III and II), but the most of the class I mutations are not predicted as

highly damaging. This outcome can be understood by looking at how the mutations are distributed around the binding site of G6PD (Figure 3.53). The most of the class III and II are located more than 10 Å away from any of the binding site, while class I mutations tend to be closer to the binding sites (< 5 Å).

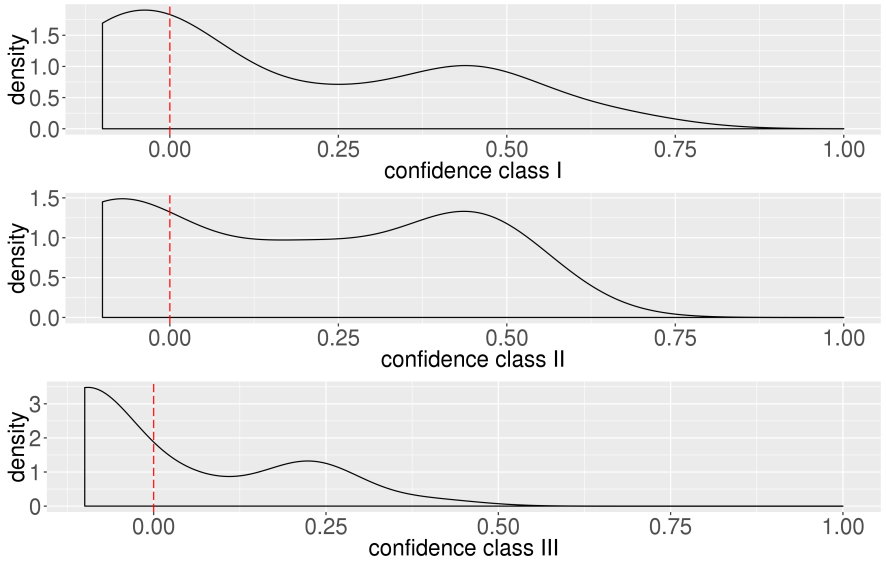


FIGURE 3.52: Distribution of the SAAPpred confidence of the mutations associated with known G6PD variants. The mutations that are existing variants, but were not predicted damaging by SAAPpred (SNP) are on the left of the dotted line (only for representation purposes their values were fixed to -0.01).

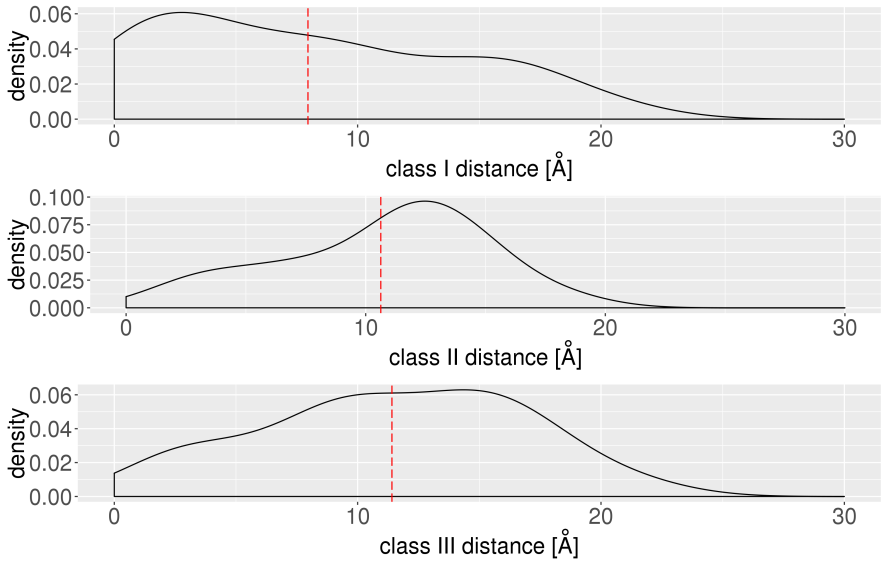


FIGURE 3.53: Distribution of the distances of the residues that are associated with known variants from any of the binding site. The dotted line indicates the average values of each class.

The intensity of the damage of a mutation could therefore correlate with its distance from a binding sites. Mutations close to the binding sites are likely to be damaging

because they alter the protein binding capability, rather than because of a serious structural damage. In contrast, for mutations far from a binding sites the nature of the substitution (e.g. G to R rather than G to A) plays a more important role in defining the resulting damage to the protein structure. Residue L128 for example, is far from any binding site and causes a class II phenotype when mutated to an arginine (L128R), and a class III when mutated to a proline (L128P). Several class I mutations are not detected as damaging by SAAPpred because very subtle changes close to a binding site can have a significant effect even though there is a minimal structural disruption. This suggests the possibility of adding distance to a binding site residue as another parameter in SAAPpred.

Overall the simulation data indicate that the mutations are relatively stable and more than globally disrupting the enzyme structure, they seem to accentuate the dynamic nature of G6PD, causing instability in the neighbouring areas. Mutant Y70H for instance, is capable of inducing a premature unfolding of the N-terminus, eventually causing a distortion in the co-enzyme binding site geometry. G204R increases the rigidity of the G6P binding site, probably affecting the way G6PD rearrange itself around G6P, ultimately affecting its binding. As a last example, the arginine found in G306R disrupts the stability of the β sheet at the C-terminus, increasing the instability and binding capability at that terminal region. Because the ultimate goal of this project was the hunt for shared modifications among mutants, it is not possible to state that an antibody-based assay could be developed to detect mutations in G6PD. It is possible to generalise the effects with similar outcomes (e.g. it destabilises the N-terminus), but apart from the A^- mutant, which presents a well defined structural alteration, all the other mutants have very subtle and different local changes in the structure.

This outcome is perhaps not surprising. Most mutations predicted as damaging with the highest confidence by SAAPpred have not been observed in people, presumably because such mutations abolish activity completely and consequently are lethal. Most mutations that cause depressed activity, and are observed in nature, tend to be predicted by SAAPpred as damaging with a lower confidence. Because of the high importance of G6PD activity, mutations that cause complete G6PD inactivity are lethal and therefore not observed. Nevertheless, even these can be accommodated through minor structural rearrangements despite the protein becoming inactive. To conclude, the depressed activity observed in several G6PD variants is likely to be connected with the disruption of key features, such as substrate binding and dimer stability, more than global or partial unfolding of G6PD.

This chapter demonstrated how the effects of the mutations act on a local scale and are not capable of causing global unfolding of G6PD structure. Because all the simulations done were in the order of nano-seconds, it was necessary to replicate and increase the confidence of these results, to reach the micro-second time scale. In an attempt to observe large scale movements and rearrangements, coarse-grained simulations were used and the work done is presented in Chapter 5.

Chapter 4

Additional studies: metadynamics

Pro172 constitutes the central residue of the conserved EKPxG peptide involved in $NADP^+$ binding, and it has been found in both *cis* and *trans* conformations. It was proposed that the cis-trans isomerisation of Pro172 plays a key role in the correct positioning of both G6P and $NADP^+$ for the catalysis [73]. In Chapter 3 this mechanism was monitored to determine the reality of its existence, and this chapter describes an attempt at using metadynamics to obtain a better understanding of the role played by Pro172 in substrate-co-enzyme interaction.

4.1 Metadynamics

Detailed all-atom MD simulations are computationally expensive. The potential energy functions used are complex, allowing only small integration steps (fs) to be used to capture the different motions in the systems. These factors limit the capacity of studying some of the most important biological phenomena such as protein folding, molecular recognition or structural transitions. Coarse-grained models (Chapter 5) are generally capable of overcoming these limitations at the expense of detail. Sometimes the sample-accuracy compromise is acceptable, but for cases where the atomistic description cannot be replaced, metadynamics can be successfully used to observe events out of reach for normal MD simulations (e.g. rare events) [140–142]. Metadynamics works by exploring the properties of a system as a function of a finite number of collective variables (CVs), which are geometric parameters that describe some chemically important process in the system. Anything can be treated as CVs, but ideally the best CVs have clear different states in the dynamics (e.g. initial, intermediate and final conformations) and should

also normally be difficult to sample, meaning that they describe slow events. Examples of CVs are atom pairs, distances, angles, dihedrals, number of contacts, hydrogen bonds, and so on. In normal dynamics the low probability of spontaneously escaping a minimum is the basis of the long time scale required to observe these rare events. The increased sampling achieved by metadynamics is the result of the addition of Gaussian functions to the potential of the selected CVs, that allows the filling of the potential well and the consequent escape from the current minimum [143, 144] (Figure 4.1).

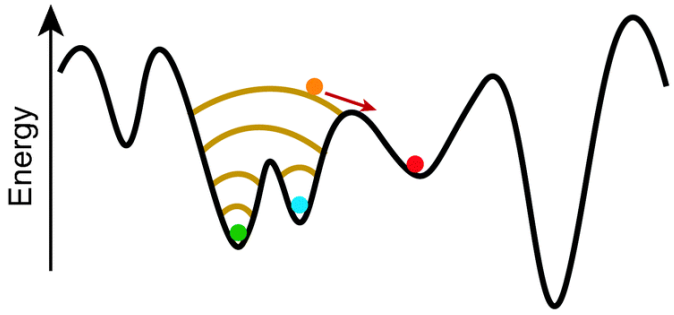


FIGURE 4.1: Schematic representation of the function mechanism of metadynamics. Gaussian functions (yellow lines) are added to the potential energy. In time, enough Gaussians will have filled the potential well and allowed the green ball to escape from the current minimum to another (red ball). [145]

Generally the external potential added is modelled as:

$$V_G(S(x),t) = \omega \sum_{t'=\tau_G,2\tau_G,\ldots} exp(-\frac{(S(x)-s(t'))^2}{2\delta s^2}) \tag{4.1}$$

where $s(t) = S(x(t))$ is the value of the CV at time t, ω is the Gaussian height, δs is the Gaussian width and τ_G is the frequency of Gaussian additions. If ω and δs are too large, the PES will be explored quickly, but large errors will be present, while if these quantities are too small, the calculation will be accurate but very slow. The time required to escape a minimum also depends on the number of Gaussians that are added to the system, and this number is proportional to $(\frac{1}{\delta s})^d$ where d is the number of CVs used. Because of this relationship, the choice of the CVs and the moments of the Gaussian must be made carefully. Parallel tempering metadynamics (PTMetaD) [146] was the method used in the project, and it differs from standard metadynamics in the fact that multiple metadynamics replicas run in parallel at different temperatures. During the simulation, exchanges of coordinates between adjacent replicas is attempted with frequency equal to $\frac{1}{\tau_x}$ (with τ_x being the frequency of Gaussian addition). If the exchange is accepted, the coordinates are swapped and the velocities rescaled. With PTMetaD, performance of both parallel tempering and metadynamics are increased and higher energy regions are explored.

In MD simulations, the temperature (T) plays a key role because it provides the system with energy (see Section 2.3.2). By increasing the temperature of the simulation box, the system acquires kinetic energy, increasing the possibility of escaping from the current minima and sampling other conformational space. The functional biological structures are stable at low potential energy, meaning that canonical MD at low temperatures tend to get trapped in local point of minimum energy. However, keeping the temperature too high is dangerous, as the conformations explored can be significantly different from the biologically functional ones. One approach is to enhance the sampling of the conformational space by running several independent replicas at different temperatures, and periodically exchange the coordinates of the system among the replicas. This allow conformations with similar potential energy to sample, over time, conformations at different temperatures, overcoming energy barriers more easily.

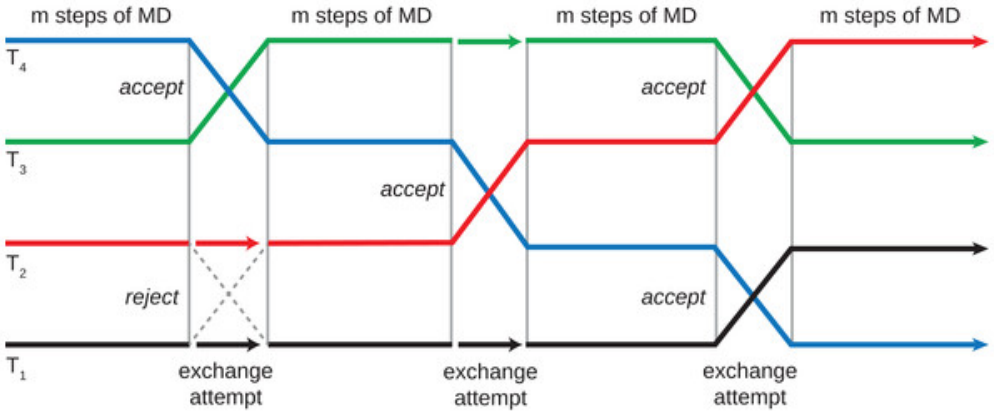


FIGURE 4.2: Schematic representation of the REMD run [147]. At certain intervals, the conformations of two neighboring replicas are exchanged based on acceptance criterion.

4.2 Metadynamics methodology

The software used was GROMACS with the PLUMED [148] plug-in being responsible for the metadynamics calculations. PLUMED groups a set of routines that work on top of MD software (e.g. GROMACS) and is designed to perform free-energy calculations as a function of many CVs using different methodologies, such as metadynamics. Only the wild-type was studied and two simulations of 16 and 6 temperatures were performed. For both, the temperature range was obtained by increasing the initial temperature (310 K) by 5 K, until the desired number of temperatures was obtained. The simulation with 6 temperatures ran for 93 ns, while the other for only 13 ns. Exchanges between temperatures were attempted every 500 steps (1 ps), and the same frequency was used for the addition of Gaussians of height = 1 and sigma = 0.35 kJ/mol (0.084 kcal). Because coordinates were exchanged, the resulting trajectories were continuous with respect to the ensemble, but not with respect to the simulation time. To rebuild each trajectory through the correct temperature space, the *demux.pl* script (included in the GROMACS distribution) was used. The CVs studied were the two distances between the C_β of Pro172 and the centre of mass (CoM) of R72 and R365. The two arginines constitute the farthest extremities of the G6P (R365) and the co-enzyme (R72) binding sites, and their movement was considered as a good indication of the movement of both the substrate and co-enzyme in relation to Pro172.

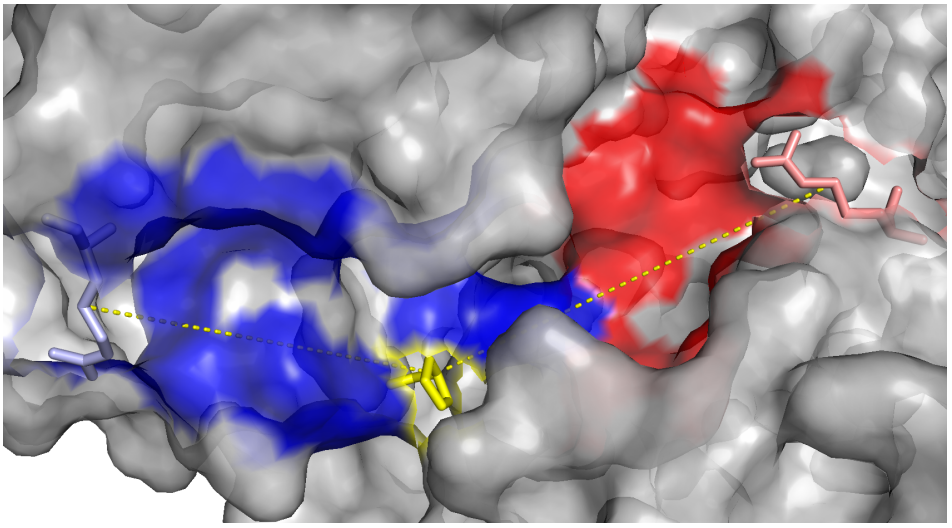


FIGURE 4.3: Representation of the distances used for the metadynamics calculation. On the surface of G6PD, the G6P and co-enzyme binding sites are coloured in red and blue respectively. At their far ends, R72 is coloured in light-blue, while R365 is in salmon. The dotted lines describe the distances between these residues and Pro172 (yellow). The figure was prepared with PyMOL.

4.3 Results

Before starting to describe the results, it must be said that the metadynamics simulations were not successful as planned. First, the time required to run a PTMetaD experiment was greatly underestimated. If a single MD simulation can make use of all the computer resources allocated, in a PTMetaD experiments, these are shared between the several replicas. On top of that, metadynamics calculations are performed at certain steps, affecting performance even further. By way of example, the 13 ns simulation with 16 temperatures took almost three months before it was discontinued. Second, the temperature interval (5 K) was too big to allow the correct exchange rate between replicas. A more correct setting, for a system as big as G6PD, would have had more than 150 replicas with temperature intervals in the order of 0.4 K. To perform such a simulation was unrealistic because of both hardware and time requirements. These two limitations were mainly the result of a combination of the size of the system and inexperience in performing metadynamics simulations. Nevertheless it was possible to obtain some data from the longest simulation.

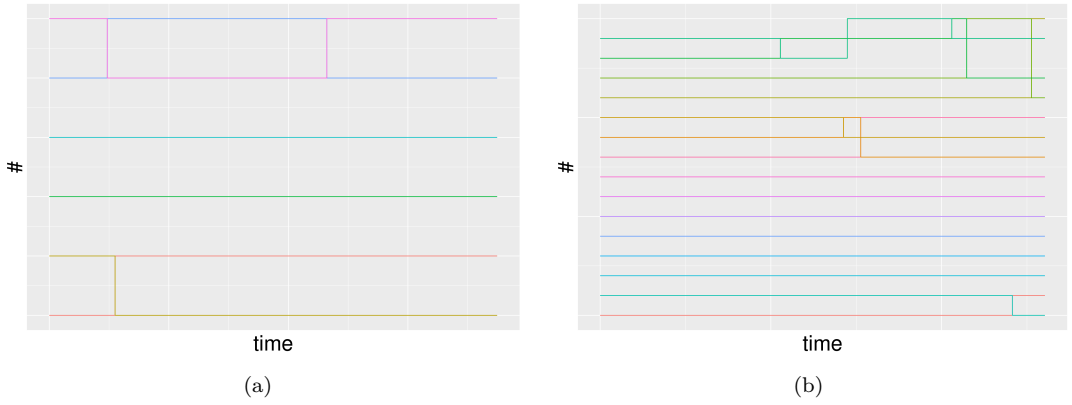


FIGURE 4.4: Diagram of the exchanges between replicas occurred in the metadynamics simulations with (a) 6 and (b) 16 temperatures.

Figure 4.4a shows that only three exchanges occurred during the dynamics. The first exchange happened after 13.7 ns between the low temperature replicas at 310 K and 315 K (Figure 4.4 in red and dark yellow), while the other two events involved the replicas at the highest temperatures, 330 K and 335 K (Figure 4.4a in blue and magenta), and happened after 12 and 58 ns respectively. The low exchange rate indicates that instead of having a PTMetaD run, the experiment can be considered as six independent metadynamics runs. When the dynamics were analysed, it was possible to recognise three behavioural groups. In the first group, the dynamics did not escape the minimum of energy and, as a consequence, the metadynamics did not fully explore the CVs' space.

In replica 1, for example, the global FES profile (Figure 4.5) shows only one single potential well and, when the FES is estimated as a function of the CVs (Figure 4.6), it appeared clear that the system was stuck in a minimum.

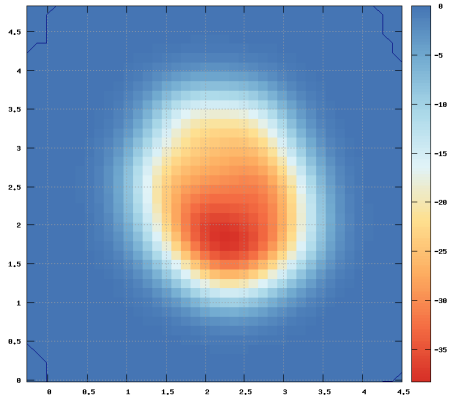
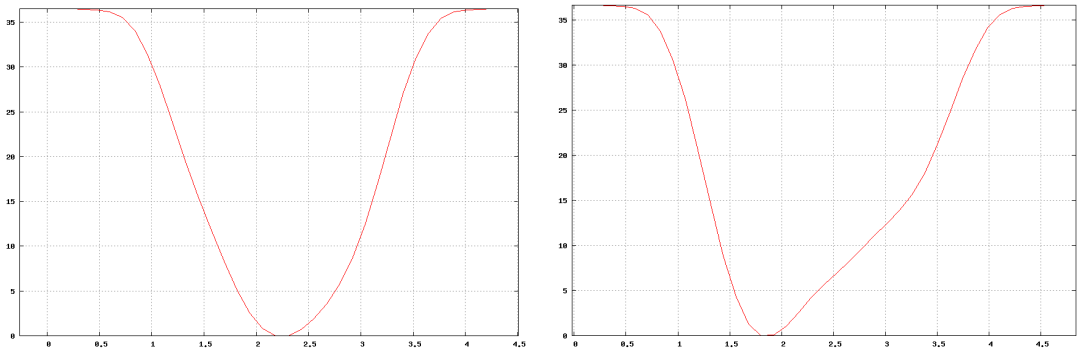


FIGURE 4.5: Example of single minima FES, obtained by 2d projection of the values of CV1 (x-axis) over CV2 (y-axis) for replica 1.



(a) (b)

FIGURE 4.6: Example of single minima FES, obtained by 1d projection of the values of (a) CV1 and (b) CV2 only, for replica 1. The x-axis describes distance values in nm while on the y-axis the values in Kcal/mol indicates the depth of the potential well.

In the second group, the dynamics had enough energy to exit the initial minimum, but it was not capable of fully sampling other minima (Figure 4.7). In replica3 for example, both CVs reached saddle points (Figure 4.8a) without visiting any other minima. The last group contains those dynamics runs that were capable of escaping from the potential well and successfully sampled other minima (Figure 4.9). It is not surprising that this group is composed of the higher temperature replicas: 330 and 335 K.

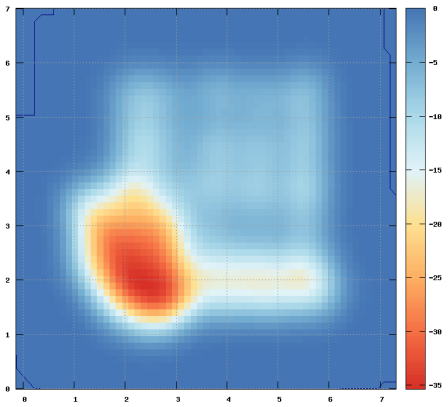
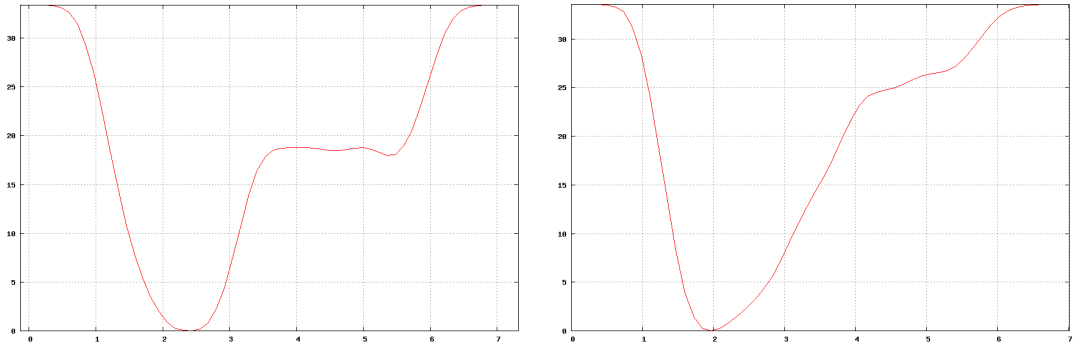


FIGURE 4.7: Example of FES with multiple minima FES, obtained by 2d projection of the values of CV1 (x-axis) over CV2 (y-axis) for replica 3.



(a) (b)

FIGURE 4.8: Example of single minima FES, obtained by 1d projection of the values of (a) CV1 and (b) CV2 only, for replica 3. The x-axis describes distance values in nm while on the y-axis the values in kcal/mol indicates the depth of the potential well.

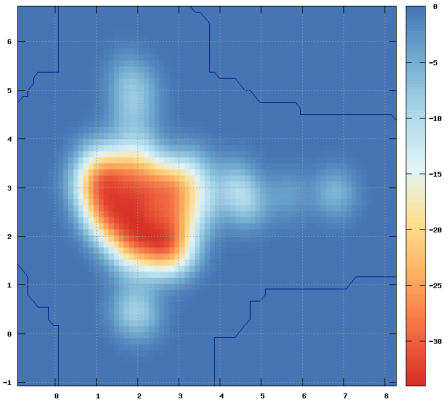
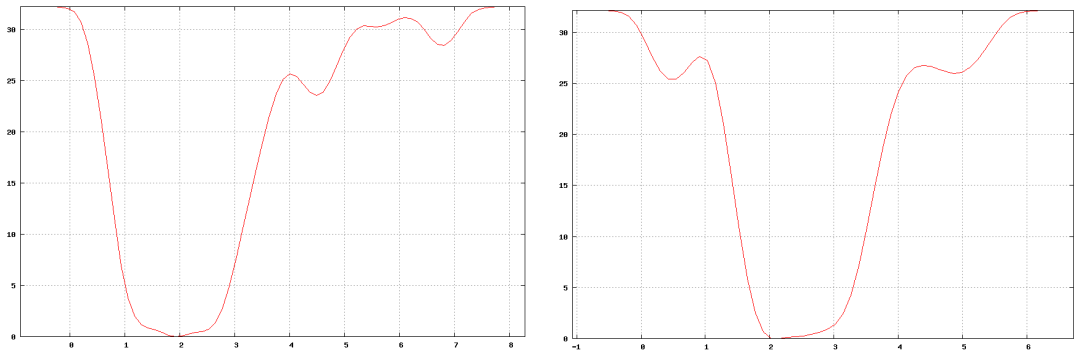


FIGURE 4.9: Example of multiple minima FES, obtained by 2d projection of the values of CV1 (x-axis) over CV2 (y-axis) for replica 5.

The proposed mechanism for Pro172 functioning is that by moving, it favours the interaction of K171 with G6P through its terminal amino group and with the co-enzyme through its carbonyl group [73]. To accomplish this role, it is fundamental for Pro172 to



(a) (b)

FIGURE 4.10: Example of multiple minima FES, obtained by 1d projection of the values of (a) CV1 and (b) CV2 only, for replica 5. The x-axis describes distance values in nm while on the y-axis the values in kcal/mol indicates the depth of the potential well.

oscillate from and towards both binding sites. The aim of the experiment was to monitor the CVs over the trajectories to identify the values of the distances that describe distinct energy-favourable conformations. Ideally, only two values (states) should exist: one that represents the residues at their maximum and the other at their minimum distance from Pro172. In the reference PDB file (2bhl), the centre of mass (CoM) of R72 and the CoM Pro172 are 20.8 Å apart while the COM of R365 is 21.1 Å from the COM of Pro172 (Figure 4.11), suggesting that the expected values of the two states should be in the neighbourhood of 21 Å (or 2.1 nm).

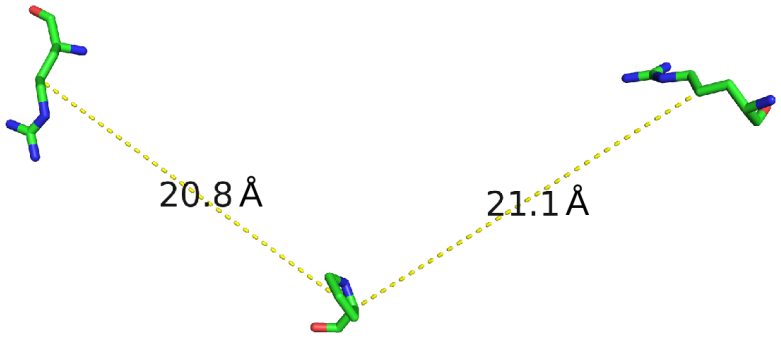
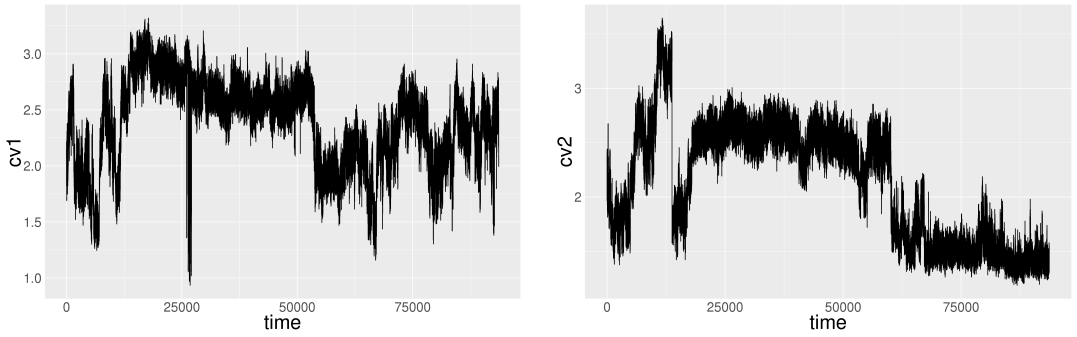


FIGURE 4.11: Measure of the value of the two CVs in the reference structure (PDB file 2bhl).

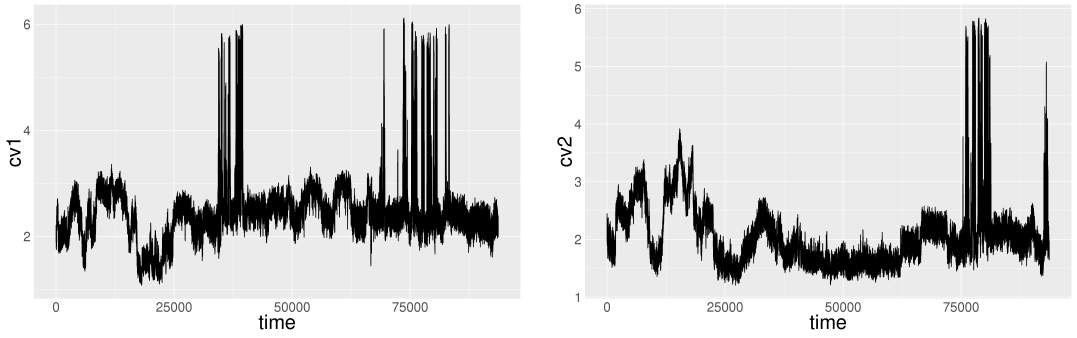
The projections of the CVs along the trajectory of one replica (Figure 4.12) indicates that while the values of CV1 only oscillated around a single state centred at 2.4 nm (Figure 4.12a), CV2 had two well defined distances; one with oscillations centred at 2.5 nm and the other at 1.5 nm. In another case, the CVs' projections only assumed one distance state each (Figure 4.13). The overscale (5 or 6 nm) peaks that are found in some replicas, suggested that the dynamics were close to overtaking some saddle point,

but the energy of the new state was too high and therefore could not be maintained.



(a) (b)

FIGURE 4.12: Projection of (a) CV1 and (b) CV2 over time for replica 1.



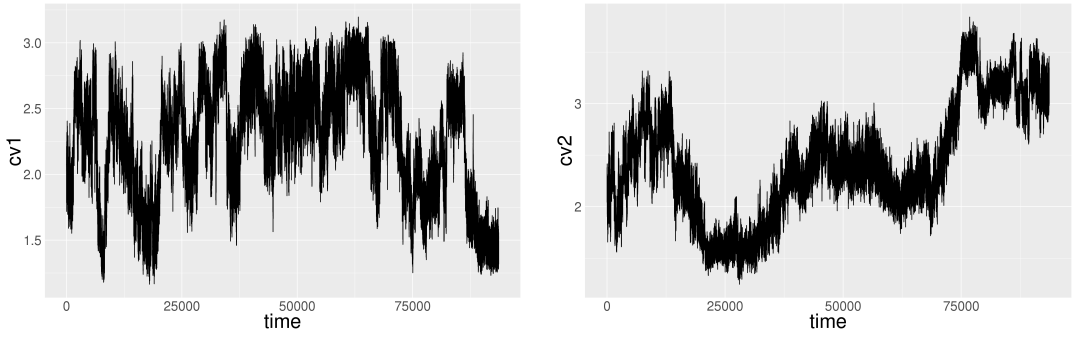
(a) (b)

FIGURE 4.13: Projection of (a) CV1 and (b) CV2 over time for replica 3.

In the most energetic replica (replica6), it was possible to notice how the decrease of one distance (Figure 4.14a) was balanced by the increase of the other (Figure 4.14b). This could be in accordance with the proposed mechanism in which the positioning of the binding sites is controlled by the movement of Pro172. When Pro172 gets closer to one site (e.g. the co-enzyme binding site), the distance between Pro172 and the other site (e.g. the G6P binding site) increases.

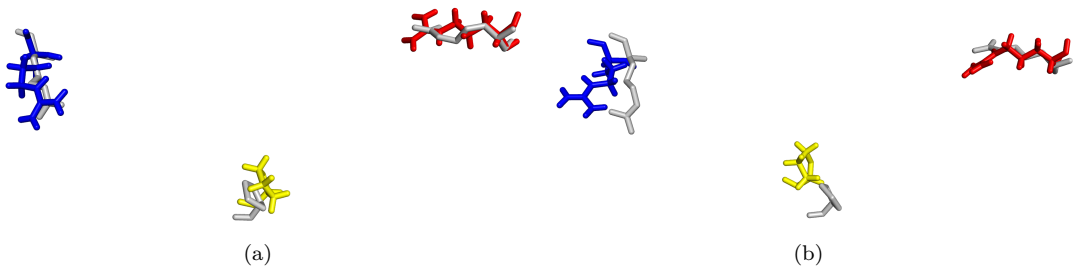
When the coordinates of the conformations with lower energy were extracted from the trajectories, a connection was noted between the movements of the two arginines and the position of Pro172. When Pro172 does not move during the trajectories (Figure 4.15) and maintains the same position as in the reference structure (2bhl), there are no changes in either position or orientation of the R72 and R365 side chains (Figure 4.15). Whereas, when Pro172 moves towards R72, R72 orients itself in the direction of Pro172 while R365 maintains its position (Figure 4.16a). This movement is enhanced at higher temperatures, where the Pro172-R72 distance is even shorter (Figure 4.16b).

These movements seems to suggest that instead of moving in unison, Pro172 gets closer



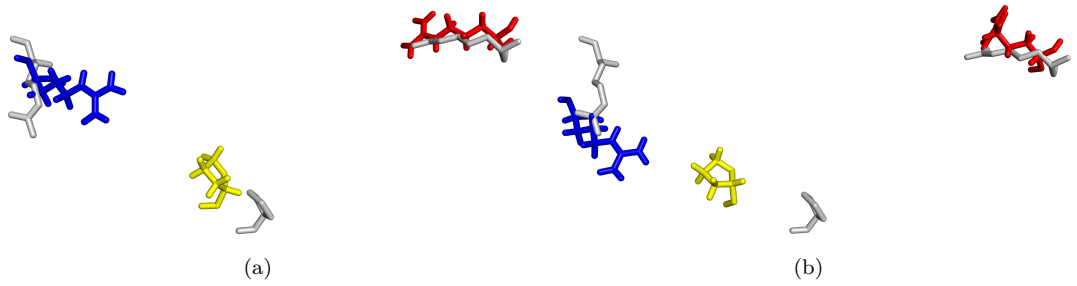
(a) (b)

FIGURE 4.14: Projection of (a) CV1 and (b) CV2 over time for replica 4.



(a) (b)

FIGURE 4.15: Snapshot of the position of R72, R365 and Pro172 at different moment of the dynamics. When Pro172 (yellow) maintains the same position as the reference structure (grey), its distances from R72 (blue) and R365 (red) do not change. In grey the same residues as are found in the reference PDB structure (2bhl).



(a) (b)

FIGURE 4.16: Snapshot of the position of R72, R365 and Pro172 at different moment of the dynamics. (a) When Pro172 (yellow) moves, it goes towards R72 (blue), which orients its side chain towards Pro172, while R365 (red) keeps its positioning stable in the G6P binding site. (b) The same mechanism is enhanced at higher temperature. In grey the same residues as are found in the reference PDB structure (2bhl).

to R72, with the effect of bringing the co-enzyme to the G6P binding site. G6PD binds G6P in its core and therefore the substrate binding site movements are more restricted than those in the co-enzyme binding site. The co-enzyme, in fact, binds in an area that is at the end of the Rossmann-like domain, which has a wider freedom of movement (Figure 4.17). As the simulations indicate, Pro172 has the role of getting the co-enzyme closer to the binding site, allowing the interaction between the two.

4.4 Discussion

Metadynamics simulations were used in order to understand the role of Pro172 in the binding sites interactions. Pro172 is the central residue of the conserved EKPxG peptide and it is directly involved in the correct positioning of the substrate and co-enzyme binding sites better. In the dimer used for the simulations (PDB code 2bhl) the proline is in *cis* conformation, but in the tetramer (PDB code 1qki), seven subunit are found in *trans*. In *Leuconostoc mesenteroides*, Pro149 (the respective of Pro172 in human) is found in *cis* conformation in all the complexes, with exception of one subunit in which is in *trans* in the absence of the coenzyme (PDB code 1dpg [149]). Because of the existence of both forms, it has been proposed that a *trans-cis* isomerisation of Pro172 could favour the movement of the αe helix, allowing Lys171 to interact with both the co-enzyme and the substrate [81]. The importance of Pro172 function is further stressed by the fact that the Volendam variant, which replaces Pro172 with a serine, exhibits class I depressed activity phenotype and its K_m values are the lowest of all the known variants [128].

The metadynamics simulations were used to monitor the distances of Pro172 to the extremes of the bindings sites, to find evidence of the oscillating movement of Pro172 from one site to the other. This movement would show how constraint Pro172 is, clarifying its role in G6PD functioning. The results indicate that the distances assume specific values in a range of 1.5 nm to 3 nm. When the conformations are extracted from the trajectories, it appears clear how these values are compatible with the residues being farthest from and closest to Pro172. Overall, the movement of the binding sites relative to the proline seems to indicate two things. First, R365, which represents the G6P binding site, is more constrained than R72 (the co-enzyme binding site) in its movement. This could be because of the substrate location in the core of the enzyme (Figure 4.17 in red). Second, the Pro172 movements are more often associated with R72 getting closer, and therefor with the co-enzyme binding site approaching the substrate. The overall conclusion is that Pro172 could really mediate the interaction between the substrate and the co-enzyme, by bringing the co-enzyme closer to the substrate in the centre of the enzyme. No clear evidence was found to support the necessity of a *trans-cis* isomerisation of Pro172, but it could be that a residue capable of adopting both *cis* and *trans* conformations leads to lower activation-energy barriers and is therefore preferred.

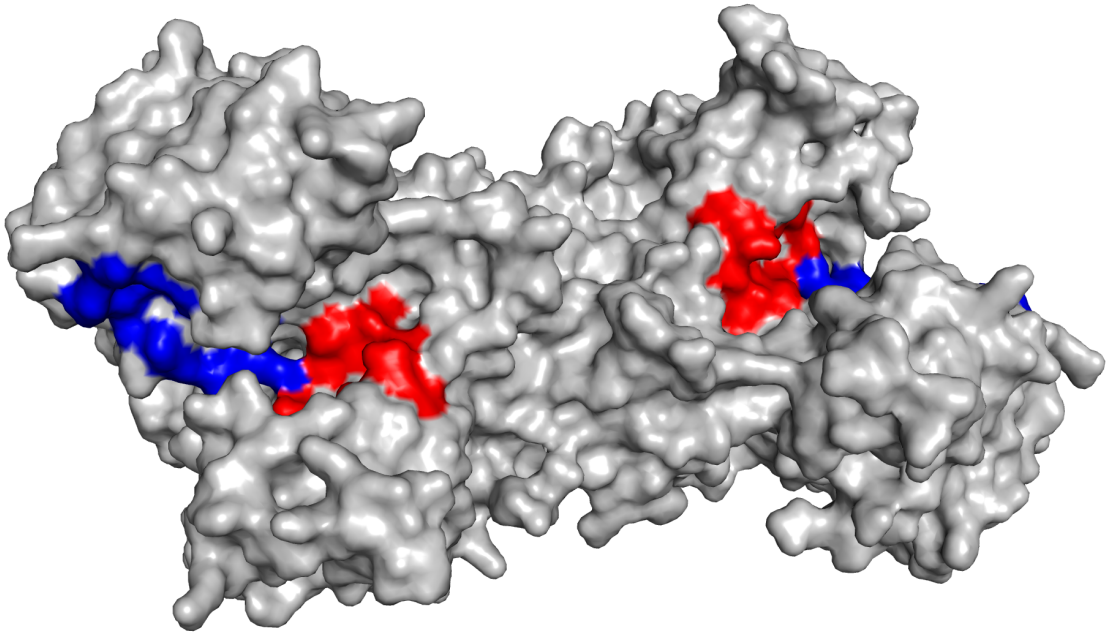


FIGURE 4.17: Location of the binding sites on the G6PD structure. The G6P binding site (red) is in the core of the enzyme, while the co-enzyme binds the blue area of the N-terminal Rossmann-like domain.

Chapter 5

Coarse-grained simulations

Several biological events happen because of the great flexibility expressed by the proteins involved. Flexibility in the structure allows proteins to move and adapt to different conditions by rearranging their structure over time. The time scale of these movements varies with the increase in the number of elements involved. Local motions happen in femto-seconds (fs), domains take micro-seconds (μs) to rearrange, and folding or unfolding can take milli-seconds (ms) or more. All-atom molecular dynamics (MD) is the most accurate computational technique to study these properties of proteins, but the great accuracy is balanced by the great computational cost. The movements of small peptides can easily be studied using hundreds or thousands of CPU/GPUs in modern supercomputers, but even these resources are not enough to characterise the dynamics of complex systems (e.g. membrane proteins) fully. Coarse-grained (CG) models are alternatives to atomistic MD simulations, which dramatically reduce the computer resources needed at the cost of detail. Because of the reduced detail, certain specific interactions (e.g. protein-solvent interactions) cannot be studied properly, but despite their simplicity, these methodologies provide high quality results.

This chapter will describe how the UNRES force field was used to study G6PD and G6PD variants' dynamics above the μs time frame. This magnitude should make the observation of large scale movements and rearrangements in the mutants possible.

5.1 The UNRES force field

The UNRES force field [150] is a physics-based united-residue force field derived as a potential of mean force (PMF), which averages the energy over the degrees of freedom that are not included in the UNRES formulation (Equation 5.1). PMF force fields are obtained from MD or MC simulations by recording how the energy changes as a function of a coordinate of the system (e.g. distance between two atoms or the variation of the torsion angles). In the UNRES force field, examples of the degrees of freedom that are averaged are the solvent and dihedral angles such as χ values. The peptide chain is represented by UNRES as a sequence of C_α atoms connected by virtual bonds to united peptide groups (p) and side chains (SC) (Figure 5.1). Only side chains and peptide groups are used during the calculation of the forces, while the C_αs are used to maintain the geometry of the other groups (dC and dX in Figure 5.1). Figure 5.2 and Figure 5.3 show G6PD represented using the classical model as a reference and as the UNRES model respectively. The energy terms in UNRES were obtained from *ab initio* quantum mechanical calculation of PMF models, while the SC terms were derived from all-atom simulations of model pairs of side chains in water [150]. The main advantage that UNRES offers is a dramatic extension of time scale up to three orders of magnitude compared with all-atom methods. This means that slow events, such as folding, which happen in μs can be observed in UNRES simulations which are only nano seconds (ns) long [150]. Because of this difference in time, all the measures of time presented in this chapter are in "molecular time units" (1 mtu = 48.9 fs) unless otherwise specified. 100 mtu of UNRES simulations are equivalent to almost 5 ps (100*48.9 fs = 4.89 ps), which corresponds to approximately 5 ns of real time simulation .

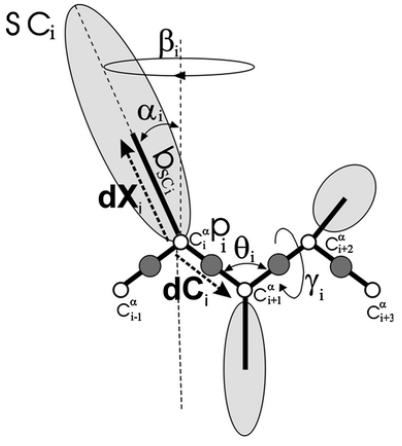


FIGURE 5.1: The UNRES model of polypeptide chain. The C_α are connected by virtual bonds with peptide-bond centres (p) and united side chains (SC) [150].

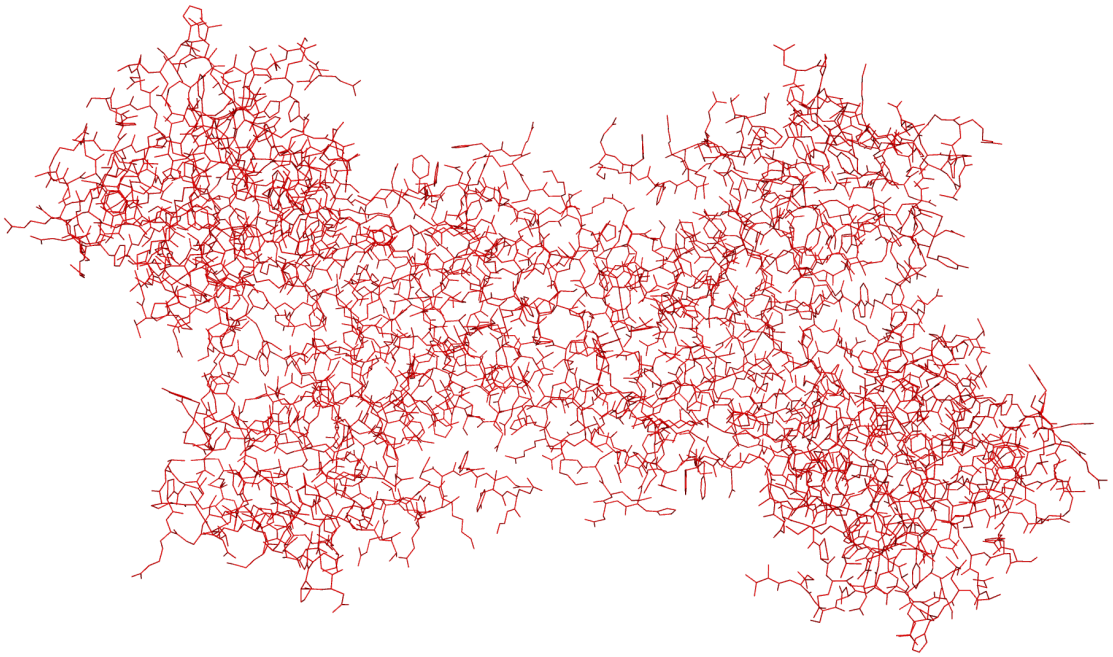


FIGURE 5.2: The all-atom G6PD dimer structure (PDB code 2bhl) visualised in line mode by PyMOL. The hydrogens were removed to improve the clarity of the figure.

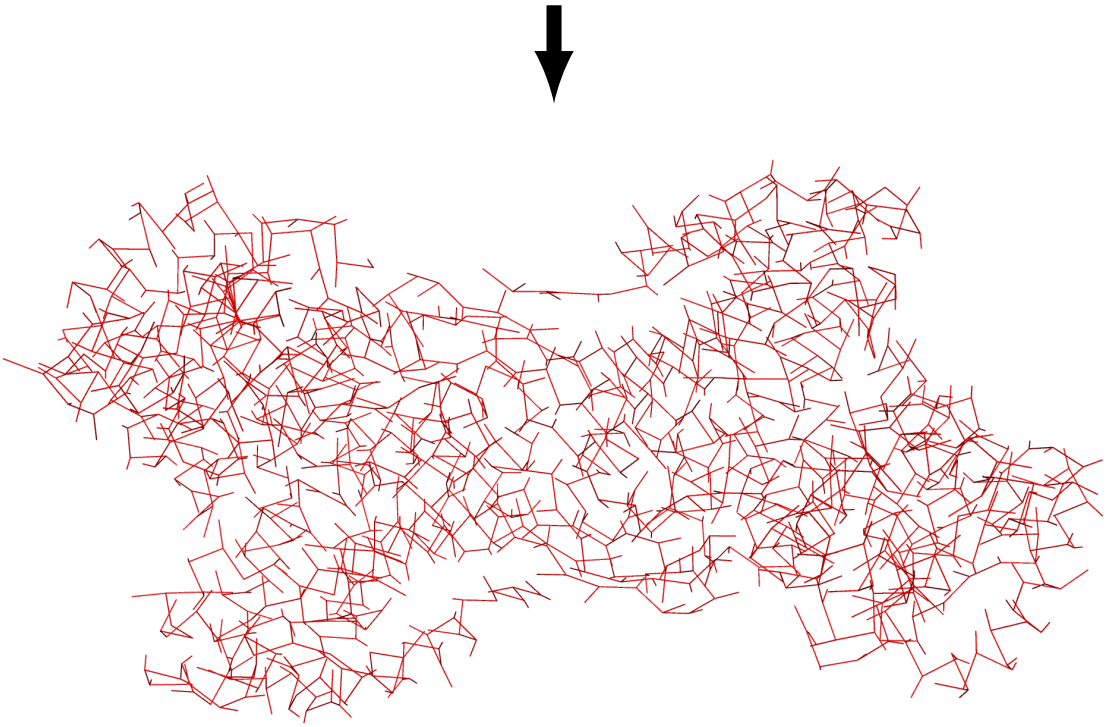


FIGURE 5.3: The G6PD dimer structure (PDB code 2bhl) after conversion to the UNRES reduced-model structure.

The equations below describe the UNRES formulation and Tables 5.1 and 5.2 give a description of the parameters and energy terms used.

$$\begin{aligned}
U =& \sum_j \sum_{i<j} U_{SC_iSC_j} + w_{SC_p} \sum_j \sum_{i\neq j} U_{SC_ip_j} + \\
&+ w_{pp}^{el} f_2(T) \sum_j \sum_{i<j-1} U_{p_ip_j}^{el} + w_{pp}^{vdW} \sum_j \sum_{i<j-1} U_{p_ip_j}^{vdW} + \\
&+ w_{tor} f_2(T) \sum_i U_{tor}(\gamma_i) + w_{tord} f_3(T) \sum_i U_{tord}(\gamma_i, \gamma_i + 1) + \\
&+ w_b \sum_i U_b(\theta_i, \gamma_i - 1, \gamma_i + 1) + w_{rot} \sum_i Urot, i + \\
&+ \sum_{m=2}^{N_{corr}} w_{corr}^{(m)} f_m(T) U_{corr}^{(m)} + \\
&+ w_{turn}^{(3)} f_3(T) U_{turn}^{(3)} + w_{turn}^{(4)} f_4(T) U_{turn}^{(4)} + w_{turn}^{(6)} f_6(T) U_{turn}^{(6)} + \\
&+ w_{bond} U_{bond}(d_i) + w_{SS} \sum_{\substack{dilsulfide \\ bonds}} U_{SS_i} + n_{SS} E_{SS}
\end{aligned} \tag{5.1}$$

where the $f_n(T)$ terms (Equation 5.2) represent the temperature-scaling multipliers, which are introduced to eliminate undesirable peaks in the heat capacity at high temperatures:

$$f_n(T) = \frac{ln[exp(1) + exp(-1)]}{ln\{exp[(T/T_0)^{n-1}] + exp[-(T/T_0)^{n-1}]\}} \tag{5.2}$$

The terms $U_{SC_iSC_j}$ and $U_{SC_ip_j}$ represent the mean free energy of the hydrophobic interactions between the side chains (implicitly containing the side chain-solvent interactions), and the excluded volume potential of the side-chain-peptide-group interactions respectively. The peptide-group interaction potential is split into two terms: $U_{p_ip_j}^{vdW}$ which describes the Lennard-Jones interaction energy between peptide-group centers, and $U_{p_ip_j}^{el}$, which is the average electrostatic energy between peptide-group dipoles. U_{tor}, U_{tord}, U_b and U_{rot} are the virtual-bond-dihedral angle terms, while $U_{bond}(d_i)$ is a harmonic potential of virtual-bond distortions, where d_i describes the length of the virtual bond. Each term is multiplied by a weight, w_x.

Similarly to every other force field, UNRES has different formulations specifically designed for the study of particular systems. The flavour used in this project is the E0LL2Y force field [151–154], a 6-12 anisotropic Gay-Berne (shifted Lennard-Jones) potential [155], particularly good in describing proteins with mixed content of α helices and β sheets.

The UNRES model was initially developed to predict protein structure by global minimisation of the potential energy using the Conformational Space Annealing (CSA) method [156, 157]. This approach was successfully used to perform *ab initio* prediction of protein structures and simulations of protein-folding phatways using the Langevin dynamics [158, 159]. The study of the thermodynamics characteristics of protein folding was further enhanced with the implementation of a replica exchange molecular dynamics (REMD) algorithm, and the support of multi-chain protein simulations [160, 161].

TABLE 5.1: The principal energy terms' weights for the E0LL2Y force field. As described in section 5.1, *SC* and *p* stand for side-chain and peptide group, while electrostatic and local cooperativity are represented as *el* and *loc* respectively.

Energy term	Value	Description
WLONG	1.00000	U(SC-SC) and U(SC,p) terms.
WSCP	1.23315	U(SC-p) term.
WELEC	0.84476	U(p-p) term.
WANG	0.62954	virtual-bond angle bending term (U_b).
WSCLOC	0.10554	side-chain rotamer term (U_{SC})
WTOR	1.84316	torsional term (U_{tor}).
WCORR4	0.00000	local-electrostatic cooperativity terms ($U^{(4)}_{el;loc}$)
WCORR5	0.00000	local-electrostatic cooperativity terms ($U^{(5)}_{el;loc}$).
WCORR6	0.00000	local-electrostatic cooperativity terms ($U^{(6)}_{el;loc}$).
WEL_LOC	0.37357	local-electrostatic cooperativity terms ($U^{(3)}_{el;loc}$).
WTURN3	1.40323	local-electrostatic cooperativity terms ($U^{(3)}_{turn}$).
WTURN4	0.64673	local-electrostatic cooperativity terms ($U^{(4)}_{turn}$).
WTURN6	0.00000	local-electrostatic cooperativity terms ($U^{(6)}_{turn}$).
CUTOFF	7.00000	cut-off on backbone-electrostatic interactions.
WCORRH	0.19212	cooperativity of hydrogen-bonding interactions term (U_{corr}).

TABLE 5.2: The force field specific parameters of the E0LL2Y force field.

Parameter	Value	Description
SIDEPAR	scinter_GAB.parm	SC-SC interaction potentials.
SCPPAR	scp.parm	SC-p interaction potential.
ELEPAR	electr_631Gdp.parm	p-p interaction potentials.
FOURIER	fourier_opt.parm.1igd_hc_iter3_3	coupling between the backbone-local and backbone-electrostatic interactions.
THETPAR	theta_abinitio.parm	virtual-bond-angle bending potentials.
ROTPAR	rotamers_AM1_aura.10022007.parm	side-chain rotamer potentials.
TORPAR	torsion_631Gdp.parm	torsional potentials.
TORDPAR	torsion_double_631Gdp.parm	double-torsional potentials.
SCCORPAR	sccor_am1_pawel.dat	torsional potentials that account for the coupling between the local backbone and local sidechain states.
BONDPAR	bond_AM1.parm	bonds.
THETPARPDB	thetaml.5parm	Additional parameters to generate random conformations
ROTPARPDB	scgauss.parm	Additional parameters to generate random conformations

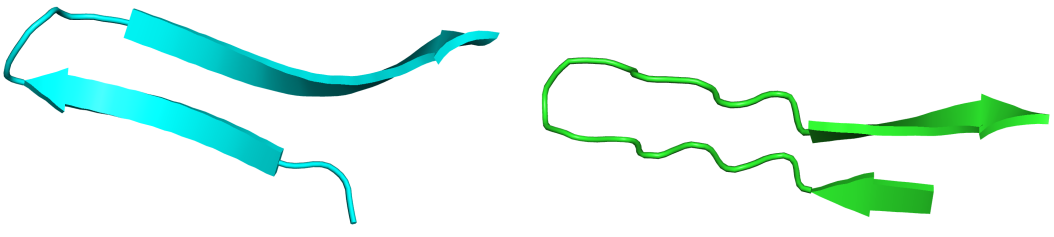
5.2 Methodology: UNRES

The UNRES force field is not implemented in GROMACS, so a completely different protocol was developed making use of the tools provided with the force field by the developers (Figure 5.5). Starting from the same structures used for the all-atom simulations, the system was first minimised in virtual-bond vectors (instead of angles) for 2000 iterations of steepest descent and then production simulations at different temperatures were started. Because of the implicit solvent adopted by UNRES, Langevin dynamics was used. Proteins exist in aqueous environments where collisions with other molecules and the solvent occur. The Langevin dynamics tries to model these forces by adding two more terms to Netwon's second law as seen in Equation 5.3.

$$\vec{F}_i - \gamma_i \vec{v} + \vec{R}(t) = m_i \vec{a}_i \tag{5.3}$$

Collisions are modelled by a random force ($\vec{R}$) and the addition of a frictional drag (γ) guarantees the correct representation of the molecule moving through the solvent (which is not explicitly included). The scaling factor of the friction coefficient used in the simulations was set to 0.01, and for each temperature, 12 independent simulations with

a time step of 1.956 fs (0.04 mtu) ran for 1×10^6 steps. Because of the different time scale between all-atom and UNRES, the 26 ns total trajectories obtained by combining all the replicas are equivalent to approximately 26 μs of all-atom simulations. Energies were saved every 2 ps and conformations every 5 ps. To control the errors of the integration method better, the Multiple Time Step (MTS) algorithm [162] with a maximum number of time-split steps of 512, was used. The MTS scheme integrates the slow-varying and the fast-varying forces of the system with different time steps. Computationally expensive forces such as electrostatic interactions, are evaluated less frequently than the fast forces, such as bond stretches, allowing a significant speed-up in the calculation. At the end of the simulations, the energy values, the rmsd and the gyration radius were directly extracted from the different outputs and the full-atom structures of the trajectories were reconstructed from the UNRES reduced model trajectories, using the script *doitGROMACS.sh* (see Appendix A). For this task, the script uses the *catomain* program, that adds backbone atoms to a C_α-only PDB file using a modified version of the method of Levitt [163]. Finally, to extract the most significative conformations, cluster analysis was performed using the Ward's minimum variance method [164]. For each cluster, the conformations were energy-ranked and only the structures with lower energy were considered as representative. Because of the limited tools provided with the UNRES force field by the developers, two main limitations affected this protocol. First the UNRES model had some distortions in the backbone representations, meaning that β strands were no longer assigned correctly by the ss-assignment software, such as dss in PyMOL (Figure 5.4). Second, to run the cluster analysis, the two chains in G6PD were linked by a 'dummy' residue inserted in the middle instead of TER records. This only affects the final PDB file (not the calculations) and was the only way of performing cluster analysis on multi-chain trajectories.



(a) (b)

FIGURE 5.4: Example of distortion in the β strands. (a) The wild-type as found in the reference PDB file (2bhl) and (b) after reconstruction from the UNRES reduced trajectory.

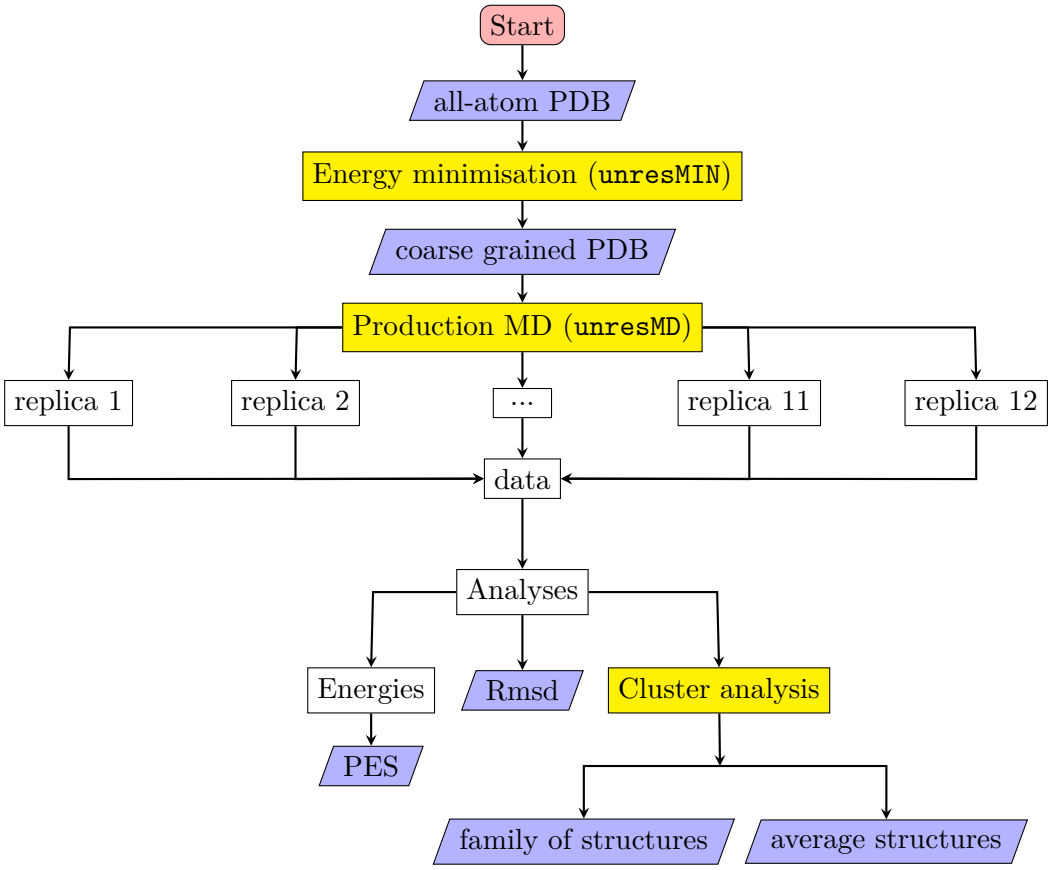


FIGURE 5.5: Diagram explaining the protocol used for the simulations and the analyses with the UNRES force field. I/O files are in blue while the UNRES tool used are in yellow. The binary 'unresMD-mult_ifort_MPI_E0LL2Y.exe' was used for the energy minimisation (`unresMIN`) and the MD steps (`unresMD`), while 'cluster_unres_ifort.exe' was used to perform the cluster analysis. Both binaries are distributed together with the force field. In-house scripts (some of which implemented in *doitGROMACS.sh*, see Appendix A) written in bash, perl and R, were used to extract and plot the data from the various files produced by the UNRES tools.

5.3 Wild-type

TABLE 5.3: The average values over the trajectories of the potential energy (Epot), temperature (K), rmsd and radius of gyration (Gyration) for the replicas of the wild-type at 310 K. For Emin, the rmsd is calculated using 2bhl.pdb (all-atom) as reference, while all the other rmsd values use the structure obtained from the energy minimisation (Emin) as reference.

Replica	Epot [Kcal/mol]	T [K]	Rmsd [Å]	Gyration [Å]
Emin	-2612	-	4.86	32.11
GB000	-2523	310.2	0.71	31.08
GB001	-2522	310.3	1.46	31.54
GB002	-2545	309.9	1.42	31.65
GB003	-2558	310.1	0.92	31.44
GB004	-2509	310.2	0.52	31.04
GB005	-2510	310.4	1.3	31.48
GB006	-2528	310.6	1.4	31.97
GB007	-2559	309.7	1.52	31.48
GB008	-2559	310.2	0.12	30.73
GB009	-2541	310.3	1.14	31.71
GB010	-2513	309.8	0.55	30.90
GB011	-2535	309.7	1.10	31.57
Mean	-2532.7	310.1	1	31.4

At the end of the energy minimisation, the minimised structure appeared to be more compact, compared with the correspondent all-atom structure (Figure 5.6). This is probably a direct consequence of the way UNRES was parametrised that tends to favour a polypeptide chain with reduced distances between the C_αs of the backbone.

5.3.1 Wild-type at 310 K

As expected, the simulations were capable of sampling a much larger section of the PES compared to the all-atom simulations. Figure 5.7 shows the energy surface explored by all the replicas combined. The PES profile was obtained as a function of the rmsd and the radius of gyration, and indicates that the most energetically favourable conformations had radius of gyration in the range from 30.5 Å to 31.5 Å, corresponding to more compact conformations (the reference PDB file has a radius of gyration equal to 36.16 Å). This confirms the propensity of the UNRES model to favour a more compact G6PD structure. With the exception of a very few cases (Figure 5.8) all the replicas converged to stable rmsd values (Figure 5.9), generally smaller than the starting conformation. The clear

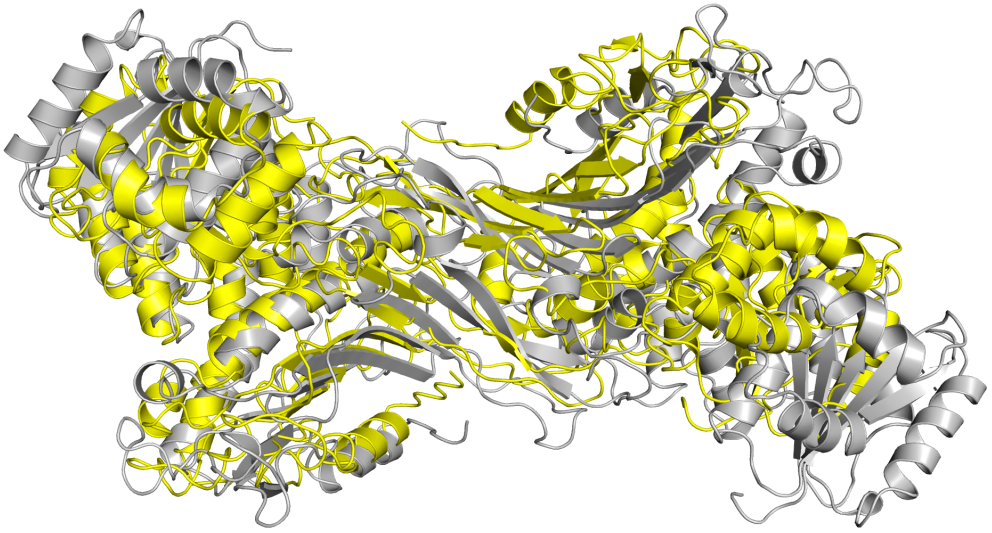


FIGURE 5.6: Superimposition of the minimised structures of the wild-type obtained with both all-atom (grey) and UNRES (yellow). The UNRES model tends to favour a more compact structure.

convergence of the simulations is another good indication of the UNRES capability of sampling the conformational space of the protein.

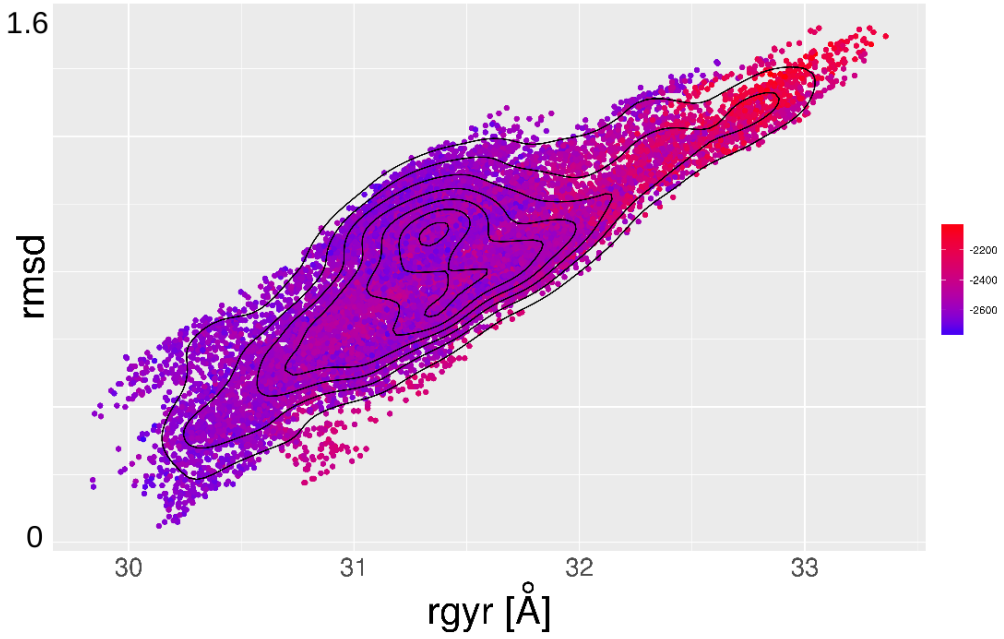


FIGURE 5.7: PES profile of all the replicas of the wild-type at 310 K combined together. The radius of gyration is indicated with 'rgyr'.

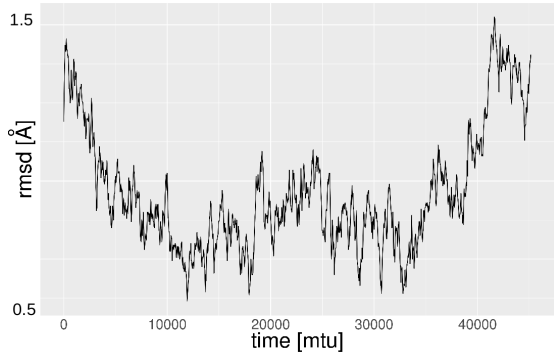
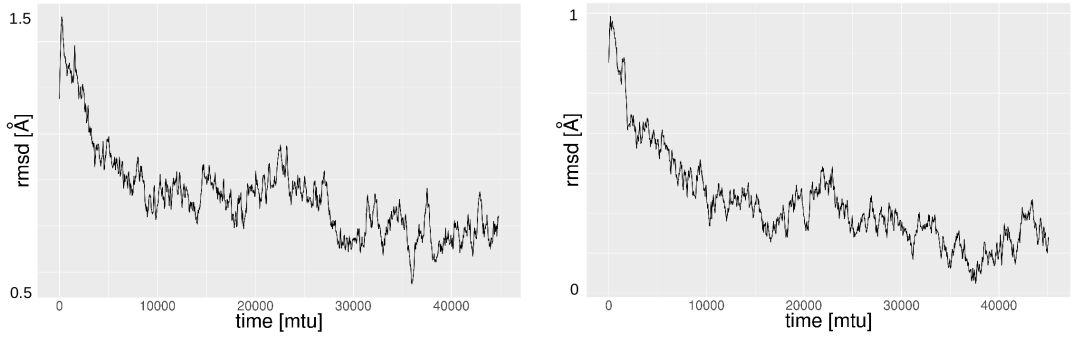


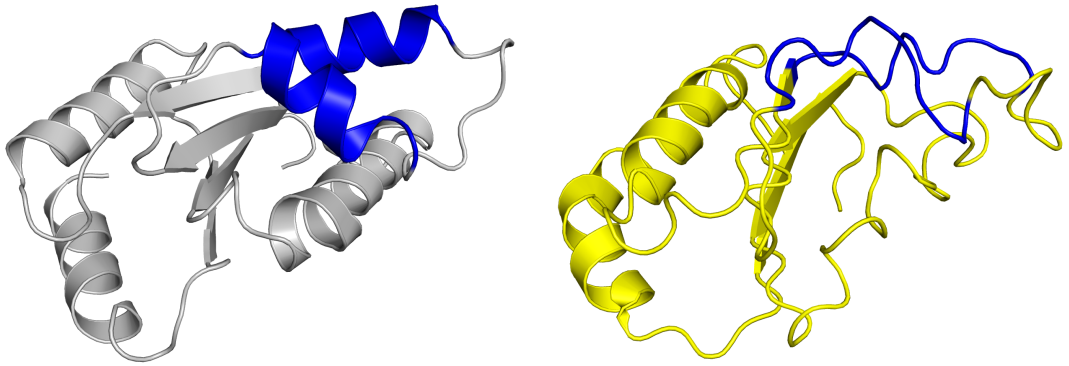
FIGURE 5.8: Example of an rmsd profile of a simulation (replica 3 of the wild-type at 310 K) that has not reached convergence. 1mtu corresponds to 48.9 fs.



(a) (b)

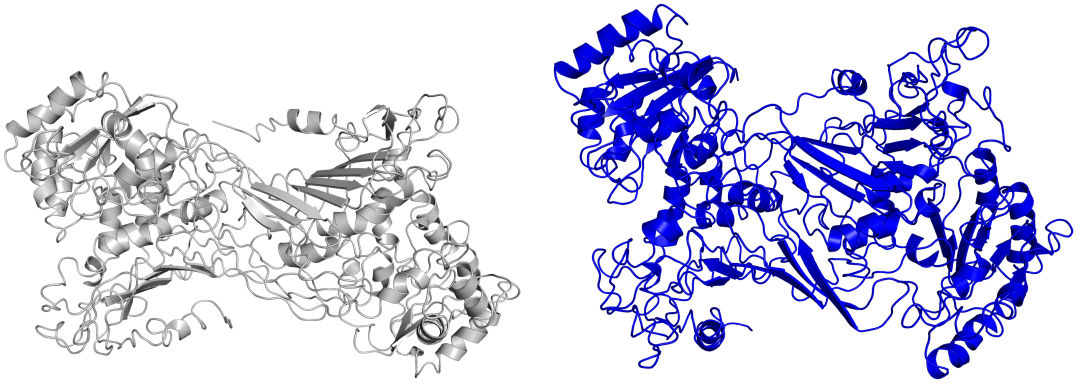
FIGURE 5.9: Examples of converged rmsd profiles of two different replicas of the wild-type at 310 K. 1mtu corresponds to 48.9 fs.

For the first 100 pico seconds of simulation, the wild-type maintained conformations that were similar to what was observed in the all-atom simulations, but here, the wild-type eventually unfolded by the end of the simulation. The central reductase domain retained a structure that resembled that of the reference PDB file (2bhl), while in the N-terminal Rossmann-like domain the unfolding was noticeable and mostly affected the helices. αa and αb, the two helices exposed to the solvent on top of the domain, unfolded completely in all the replicas (Figure 5.10), while the other helices of the domain relocated themselves in the region, causing the more compact structure indicated by the radius of gyration (Figure 5.11b). The more energetic replicas were characterised by the partial unfolding of the C-terminus region, where the helices and the last two strands (βH and βO) were replaced by coils (Figure 5.12). Overall, the results obtained indicate that the helices and the terminal regions are the weak points of the G6PD structure, and their movements cause G6PD to deform and form a more compact structure.



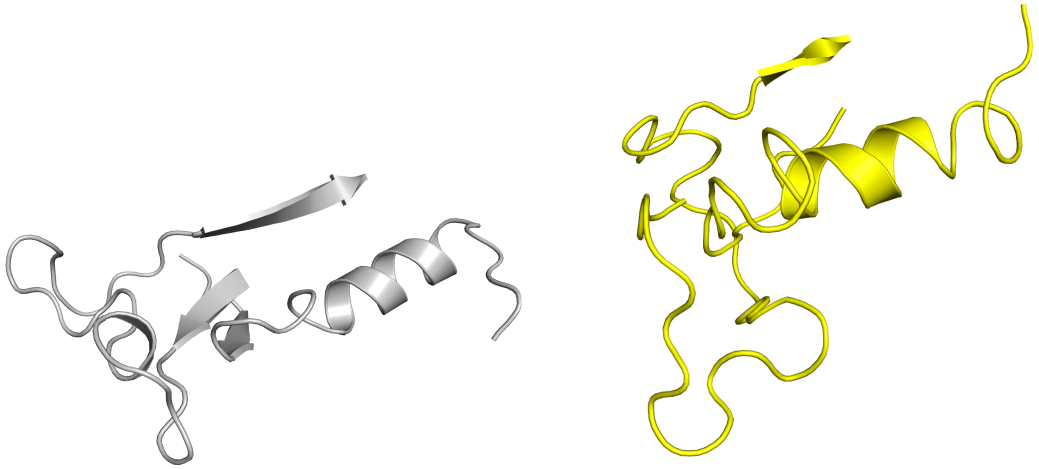
(a) (b)

FIGURE 5.10: (a) Conformation of the αa and the αb helices (blue) in the wild-type. (b) During several simulations at 310 K, both helices completely unfolded by the end of the simulations.



(a) (b)

FIGURE 5.11: (a) wild-type structure as found in the reference PDB structure, and (b) the unfolded form at the end of a representative replica of the UNRES simulation.



(a) (b)

FIGURE 5.12: (a) Folded C-terminus region of the wild-type structure as found in the reference PDB structure, and (b) the same region unfolded at the end of a representative replica of UNRES simulation.

5.3.2 Wild-type at other temperatures

To maintain a parallelism with the work done in Chapter 3, and to assess the overall performance of the UNRES model, the wild-type was also studied at different temperatures: 500, 400 and 330 K. At all temperatures, G6PD completely unfolded to the point that any type of structure was quickly completely lost (Figure 5.13). Because the aim of the project was the detection of changes in the equilibrium between the wild-type and the mutants, all these temperatures were considered to be too high to extract any useful information about G6PD stability and dynamics, and 310 K was selected as the reference temperature for the simulations with the UNRES model.

FIGURE 5.13: The complete unfolded structure of G6PD at the end of a simulation at 500 K.

5.3.3 wild-type summary

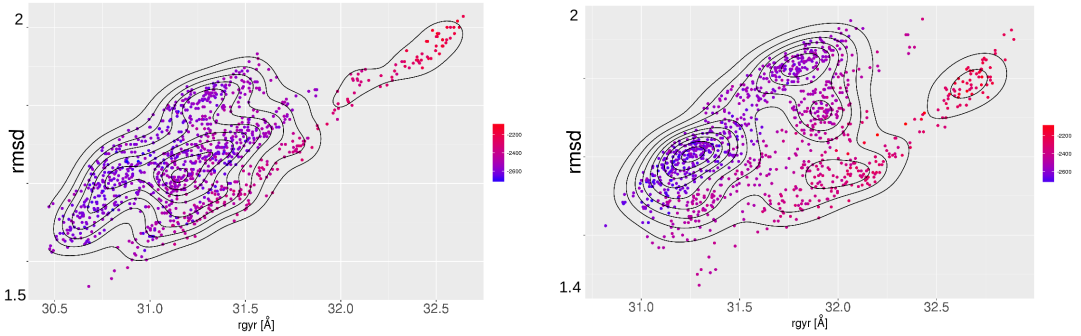
Thanks to the simplified UNRES model it was possible to increase greatly the sampling and observe the unfolding of G6PD at room temperature. Similar to the all-atom simulations, the collected data suggest that the terminal regions and the helices of the enzyme are more prone to unfolding than the central β-reductase domain. The wild-type simulations have also shown some limitations of the UNRES model. First, the united-group model is probably not detailed enough to describe the very local interactions and effects of one single residue substitution. Second, the detection of common unfolding patterns and behaviours between replicas may be difficult to detect correctly. However, the comparison with native G6PD with the effects of the mutations should probably be

possible in terms of changes in energies and in the overall stability of the structures. Expectations are that the UNRES simulations of the G6PD mutants should demonstrate that the mutants destabilise G6PD structure, by inducing a premature unfolding of the regions close to the mutation.

5.4 Damaging mutants

5.4.1 G204R

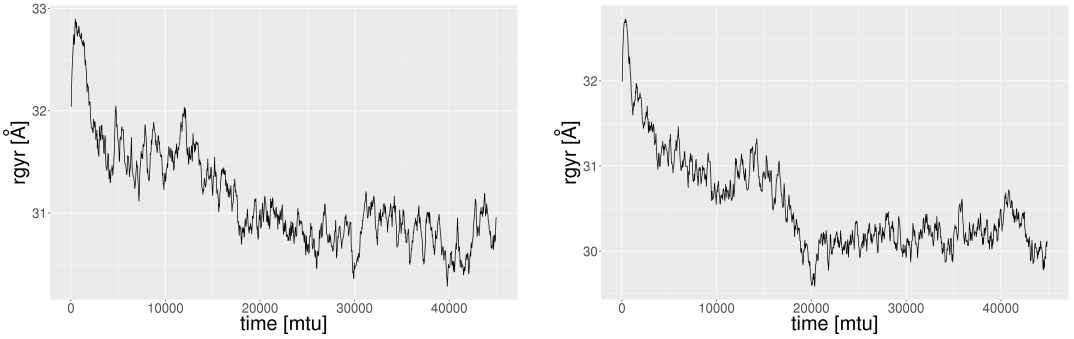
The replacement of glycine with a charged hydrophilic arginine at position 204 (R204′) was predicted as damaging by SAAPpred with a confidence of 0.8. The total potential energy is, on average, 21 Kcal/mol higher than the values recorded for the wild-type and, similarly, the average rmsd values are also higher (almost double) that of the wild-type (Tables 5.3 and 5.4). On the contrary, even if not dramatically different, the radius of gyration is slightly smaller (31.2 Å *vs.* 31.4 Å). It is possible that these values are the result of some instability effects caused by the presence of R204′. Similarly to the wild-type, all the replicas have sampled several minima and, compared with the wild-type, these minima are more distant from each other (Figure 5.14), giving another confirmation of the high sampling capability of the UNRES model.



(a) (b)

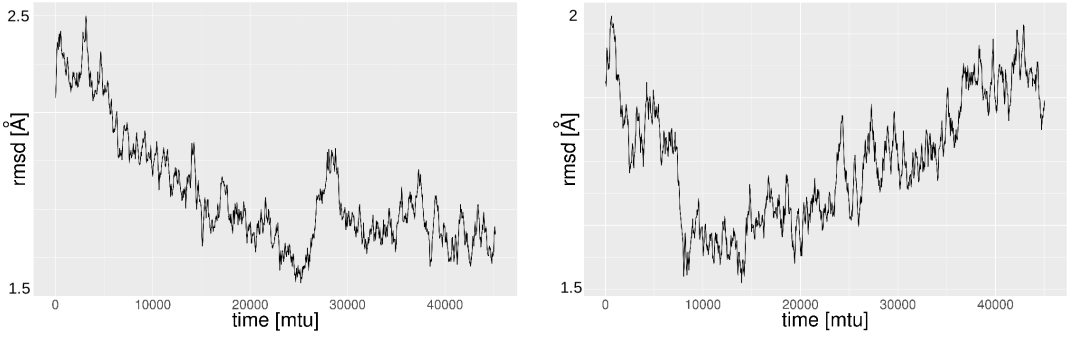
FIGURE 5.14: PES sections explored by two replicas of G204R at 310 K. The radius of gyration is indicated with 'rgyr'.

In all the replicas, the radius of gyration decreased over time, but it eventually stabilised around the values found in Table 5.4, and convergence was reached by the end of the simulations (Figure 5.15). A similar behaviour was observed for the rmsd in all the replicas (Figure 5.16a), with only few cases (replica 4 and 7) in which the rmsd increased (Figure 5.16b). Not surprisingly, these cases are observed in the replicas which have energy values above the average.



(a) (b)

FIGURE 5.15: Radius of gyration of two different replicas of G204R at 310 K. In both cases it decreases over time, but it eventually converges. 1mtu corresponds to 48.9 fs.

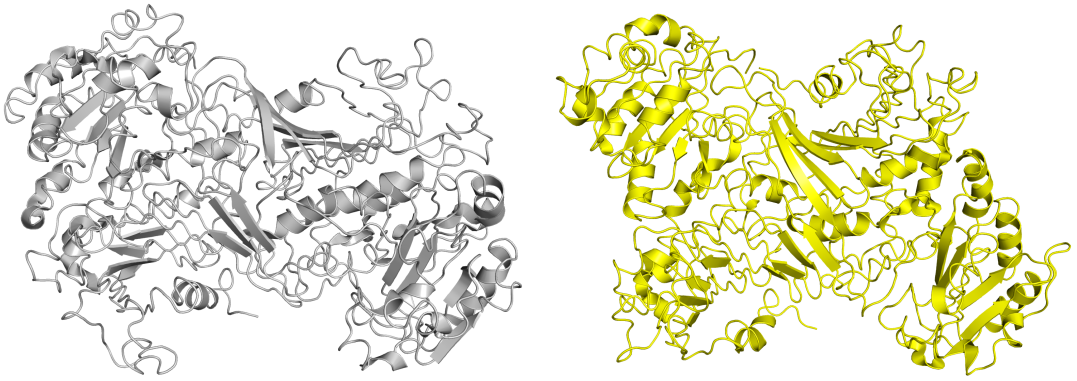


(a) (b)

FIGURE 5.16: Rmsd of two different replicas of G204R at 310 K. (a) Similar to the radius of gyration, the rmsd profile reached convergence in all the replicas, but for a few cases as shown in (b). 1mtu corresponds to 48.9 fs.

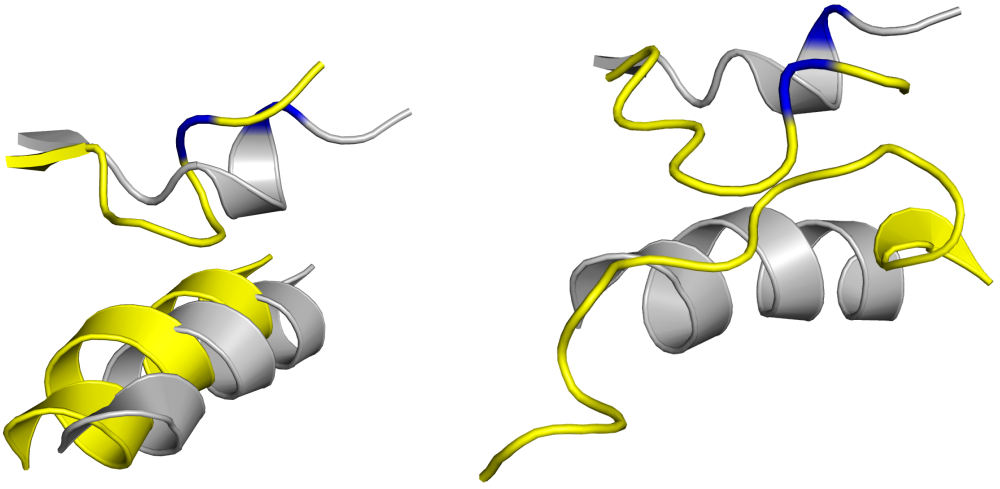
Overall G204R behaves similarly to the wild-type. The termini of the enzyme are the most unstable regions and they lead the unfolding in all the replicas (Figure 5.17). Because the structures had to be reconstructed from a reduced-detail model, it was difficult to find a clear involvement of R204′ in the unfolding dynamics. However, in both the wild-type and G204R simulations, the UNRES model broke the helix spanning residues H201-K205 between residues 204 and 205, where R204′ lies, and which provides G6P binding residues (Figure 5.18a). However in G204R that segment is less constrained than in the wild-type. The result is that this segment (R204′-K205) can move closer to the αi helix, causing its unfolding (Figure 5.18b). The αi helix constitutes the base of the G6P binding site, and was also seen in the wild-type to unfold partially, but the complete unfolding of the helix (αi) can only be seen in G204R when the segment covers αi. In the all-atom simulations, this helix was never influenced by R204′. Overall, the simulations indicates that R204′ destabilises of the surrounding area by forcing rearrangements that eventually cause the premature unfolding of the region. Because H201, Y202 and K205

are actively involved in G6P binding, any movements caused by R204′ could influence the G6P binding. What the simulations suggest is that R204′ induces a premature unfolding of the helix (H201-K205) which is now free to move around. These movements increase the overall instability of the region and, for example, cause the premature unfolding of the αi helix at the base of the G6P binding site.



(a) (b)

FIGURE 5.17: Final configuration of the trajectories of (a) replica 12 and (b) replica 11 of G204R at 310 K. The high instability of the termini of the enzyme caused G6PD to unfold and contract.



(a) (b)

FIGURE 5.18: Segment H201-K205 and the αi helix in the wild-type (grey) and in G204R (yellow), with R204′ coloured in blue. (a) The presence of R204′ causes the segment H201-K205 to unfold and to (b) move, during the simulations, inducing the unfolding of the αi helix.

TABLE 5.4: Average values over the trajectories of the potential energy (Epot), temperature (K), rmsd and radius of gyration (Gyration) for all the independent simulations of G204R at 310 K. For Emin the rmsd is calculated using the all-atom structure as the reference, while all the other values use the structure obtained from the energy minimisation (Emin) as the reference.

Replica	Epot [Kcal/mol]	T [K]	Rmsd [Å]	Gyration [Å]
Emin	-2518.7	-	5.1	31.7
GB000	-2522	309.9	1.91	31.2
GB001	-2528	310.3	1.8	31.2
GB002	-2527	309.9	1.92	31.2
GB003	-2490	310	1.8	31.5
GB004	-2535	310.1	1.89	31.4
GB005	-2493	310.1	1.77	31.2
GB006	-2474	310	1.97	31.2
GB007	-2508	309.9	1.7	31.7
GB008	-2515	309.8	1.7	30.9
GB009	-2524	310.2	1.92	30.8
GB010	-2510	310.2	2.11	31.6
GB011	-2518	309.7	1.86	30.5
Mean	-2512	310	1.86	31.2

5.4.2 Residue 306: G306R and G306S

Position 306 is located in the βH strand of the central $\alpha+\beta$ 2-layer sandwich domain, and two mutants of this residue were studied: G306R and G306S. G306R was predicted to be damaging with high confidence (0.8), but it is not an existing variant. G306S was predicted to be damaging with a lower confidence (0.64) and is known to be a class II G6PD variant. The study of both mutants gave not only the possibility of studying some predicted damaging mutants, but also the opportunity of studying the damaging mechanisms in an existing G6PD variant. In the all-atom simulations, both mutants increased the instability of the region, with a propensity to cause the premature unfolding of the surrounding β strands.

The energies calculated over the UNRES trajectories, indicate that G306R is globally less stable than the wild-type. The potential energy is almost 30 kcal/mol higher than the wild-type and this instability is reflected in rmsd values close to double the values of the wild-type for all the replicas (Table 5.5). In the simulations, G306R showed a behaviour very similar to that observed in the all-atom simulations; the presence of the arginine causes the premature unfolding of the βH and βO strands. The βH strand is where the mutation lies, and it was never found completely folded in the simulations (Figure 5.19). Instead, the βO strand contains I480 which was found interacting with the R306′side chain in the all-atom simulations (Section 3.6). The two strands are always close to each other and they reduce their lengths during all dynamics simulations (Figure 5.20).

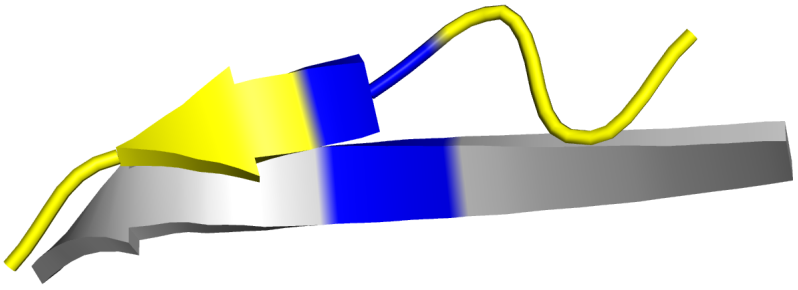
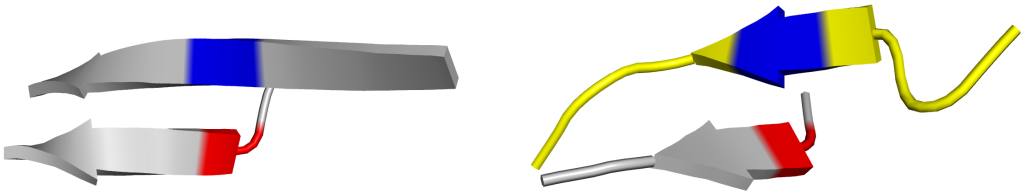


FIGURE 5.19: The βH strand in both the wild-type (grey) and G306R (yellow). Residue 306 is in blue.

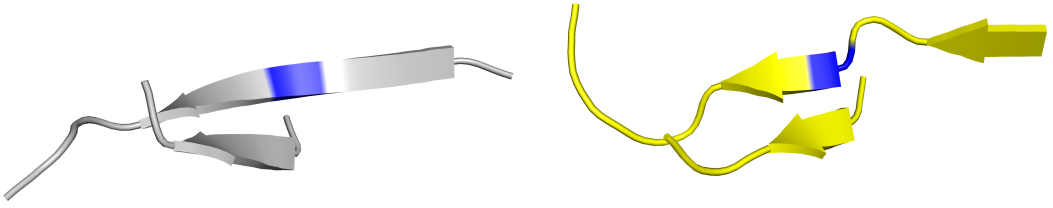
Overall the C-terminal region of G306R is more unstable than the wild-type. In the wild-type, the β strands are distorted but maintained, while in G306R the area is more disordered. However, the unfolding is not dramatically faster in G306R, possibly because of the compactness of the area which constrains the movements of the unfolding regions.



(a) (b)

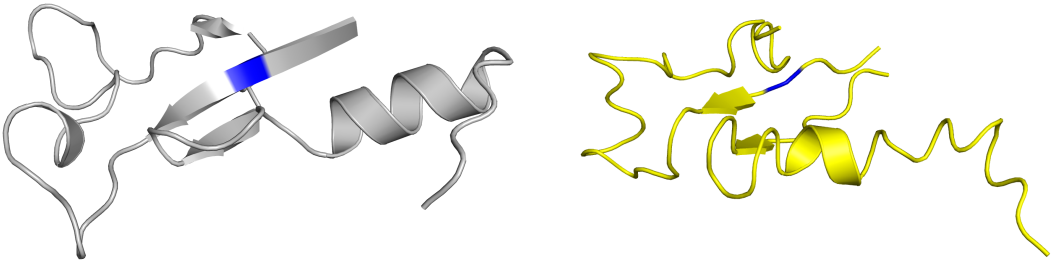
FIGURE 5.20: The βH and βO strands in (a) the wt and (b) G306R. The interaction between R306′ (blue) and I480 (red) caused βO to shrink.

G306S presents a similar mechanism. The βH strand breaks at the mutation site (Figure 5.21b), and the βO strand shortened as a consequence of the presence of the serine (S306′). The C-terminus of the protein moves more than it does in both the wild-type and G306R, and this is probably because the smaller size of S306′ allows a greater freedom of movement to the residue compared with R306′ which is kept in place. The result is a structure with a distorted structure at the C-terminal end (Figure 5.22 and Figure 5.23).



(a) (b)

FIGURE 5.21: The βH and βO strands in (a) the wt and (b) G306S. Contrary to G306R, S306′ cannot interact with I480 in βO and the damage only affect the βH strand. S306′ is in blue.



(a) (b)

FIGURE 5.22: C-terminal region of (a) the wild-type and (b) G306S, with S306′ coloured in blue.

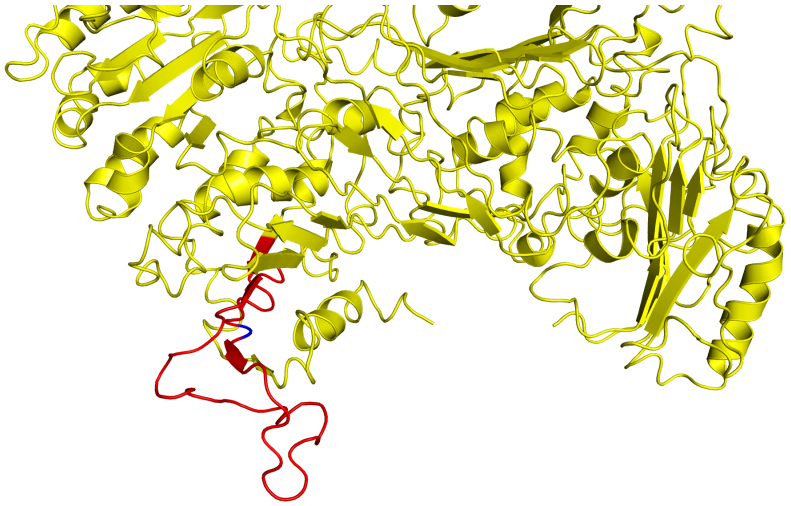


FIGURE 5.23: C-terminus of G306S at the end of a simulation, with the region affected by the mutation coloured in red.

The simulations with the UNRES model confirm that both G306R and G306S have local effects on the surroundings. They both destabilise the area, shortening the βH and βO strands and increasing the overall movement of the C-terminus. R306′, being more damaging, has a stronger effect in unfolding the strands, while S306′ seems capable of increasing the movements of the unfolding regions in the area. Residue 306 is close to the structural $NADPH^+$ binding site, suggesting a possible and direct involvement of the mutations in $NADPH^+$ stability.

TABLE 5.5: Average values over the trajectories of the potential energy (Epot), temperature (K), rmsd and radius of gyration (Gyration) for all the independent simulations of G306R at 310 K. For Emin the rmsd is calculated using the all-atom structure as the reference, while all the other values use the structure obtained from the energy minimisation (Emin) as the reference.

Replica	Epot [Kcal/mol]	T [K]	Rmsd [Å]	Gyration [Å]
Emin	-2618.7	-	5.1	31.7
GB000	-2512	309.9	2.03	30.4
GB001	-2483	310.3	1.97	30.9
GB002	-2496	310	2.2	31.5
GB003	-2510	310	1.9	31.2
GB004	-2509	310	1.9	31.8
GB005	-2497	310.4	1.8	30.9
GB006	-2553	309.9	1.9	31.2
GB007	-2488	309.9	1.7	31.5
GB008	-2517	310	1.8	31.5
GB009	-2512	310	1.9	30.8
GB010	-2506	310.3	1.95	30.9
GB011	-2498	310.4	1.8	31.5
Mean	-2506	310	1.9	31.1

TABLE 5.6: Average values over the trajectories of the potential energy (Epot), temperature (K), rmsd and radius of gyration (Gyration) for all the independent simulations of G306S at 310 K. For Emin the rmsd is calculated using the all-atom structure as the reference, while all the other values use the structure obtained from the energy minimisation (Emin) as the reference.

Replica	Epot [Kcal/mol]	T [K]	Rmsd [Å]	Gyration [Å]
Emin	-2634	-	5.1	31.7
GB000	-2538	310.2	1.8	30.6
GB001	-2543	310.3	1.7	31
GB002	-2527	310.2	1.9	31.4
GB003	-2520	310	2.1	30.8
GB004	-2510	310.6	1.4	31.3
GB005	-2532	310.2	1.6	30.6
GB006	-2524	310.5	1.5	31.6
GB007	-2529	310	1.4	31.4
GB008	-2544	310	1.7	31.2
GB009	-2506	310	1.8	31
GB010	-2543	310	1.5	31.4
GB011	-2523	310	1.8	31
Mean	-2528	310	1.7	31.1

5.4.3 A^-

A^- is the only multiple missense mutation considered in the study. It is the most common variant in Africa and in African ancestry populations and it combines V68M and N126D, resulting in a class II variant. Individually, only V68M was predicted as damaging by SAAPpred (very low confidence of 0.07), while N126D was not listed as PD (SNP confidence of 0.5). The mutant is characterised by dynamics with similar potential energy and radius of gyration, but a higher rmsd (Table 5.7) compared with the wild-type. The PES profile of all the replicas of the wild-type (Figure 5.7) was centred around a radius of gyration value of 31.5 Å, while the A^- profile is shifted toward values of the radius of gyration at the extremes of the PES explored: 30.5 Å and 32 Å (Figure 5.24).

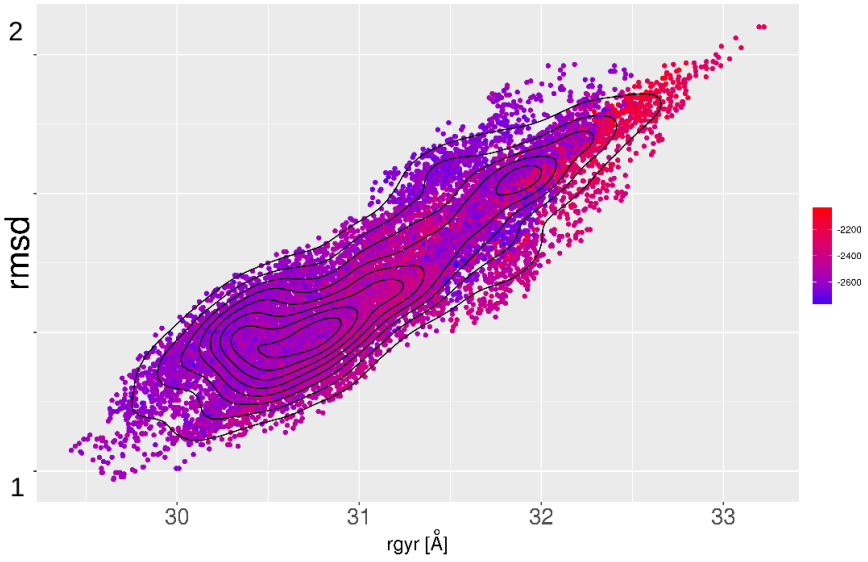
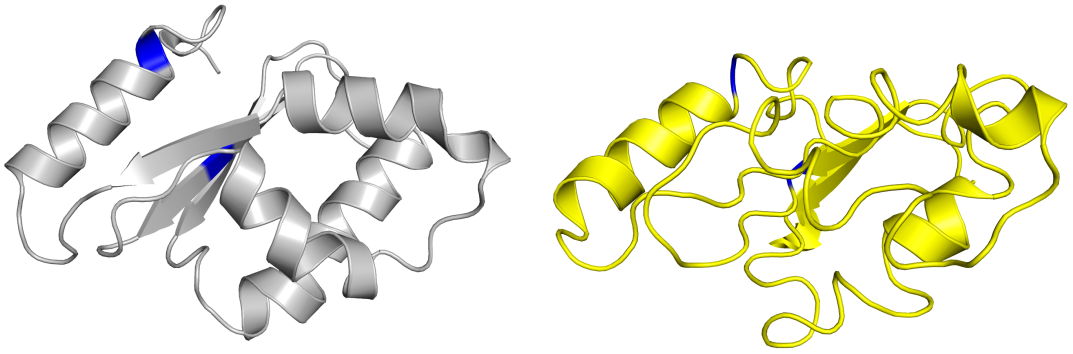


FIGURE 5.24: PES profile of all the replicas of the A^- mutant at 310 K combined together. The radius of gyration is indicated with 'rgyr'.

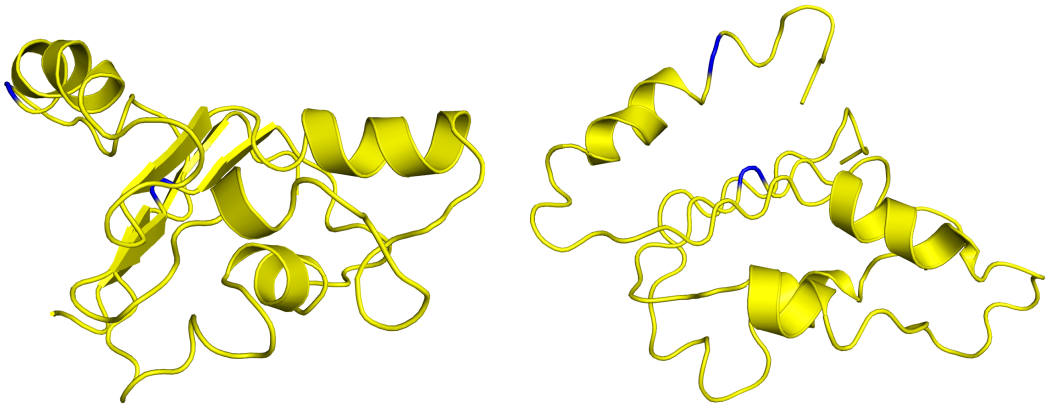
These shifts are the result of a greater instability of the N-terminal Rossmann-like domain, which is the area where the mutations are located. This domain unfolds similarly to the wild-type, but in A^- , the unfolding is faster and has more dramatic effects on the enzyme structure. It is proposed that M68′ causes shortening of the βB strand, while D126′ induces unravelling of the αc helix, resulting in the almost complete unfolding of the two external helices of the domain: αa and αb (Figure 5.25b). The instability of the region is such that the unravelling of the αb helix and of the nearby αc helix is projected to the outside of G6PD with a movement never seen in the wild-type or any other mutant (Figure 5.26b). In the all-atom simulation, the presence of M68′ and D126′ guided the rearrangement of the structures in the area, with the αc helix and the βB



(a) (b)

FIGURE 5.25: Unfolding of the N-terminal Rossmann-like domain area close to the mutations in (a) the wild-type and (b) in A^- . M68′ and D126′ are in blue.

strand which fold/unfold depending on the orientation of their side-chains (Section 3.7). In the UNRES simulations, it seems possible, in some replicas, to observe this unfolding pattern (Figure 5.26b), but the loss of details and the great instability of the area make the detection of the correct mechanisms not possible. Overall the UNRES model confirms the destabilising effects of M68′ and D126′ on the N-terminal Rossmann-like domain that accelerates the unfolding of G6PD N-terminal end.



(a) (b)

FIGURE 5.26: Unfolding of the top of the N-terminal Rossmann-like domainin A^- . (a) The twist and movement of the αc helix to the outside of the protein. (b) The unravelling of the αc helix in relationship to the βB strand. The mutations are coloured in blue.

TABLE 5.7: Average values over the trajectories of the potential energy (Epot), temperature (K), rmsd and radius of gyration (Gyration) for all the independent simulations of A^- at 310 K. For Emin the rmsd is calculated using the all-atom structure as the reference, while all the other values use the structure obtained from the energy minimisation (Emin) as the reference.

Replica	Epot [Kcal/mol]	T [K]	Rmsd [Å]	Gyration [Å]
Emin	-2689	-		
GB000	-2536	310	1.94	31.6
GB001	-2520	310.3	1.76	30.8
GB002	-2545	310	1.8	30.3
GB003	-2541	310.3	1.69	30.7
GB004	-2566	310.3	1.76	30.7
GB005	-2559	309.9	1.78	31.5
GB006	-2528	310.1	1.9	31.4
GB007	-2538	310.1	1.79	31
GB008	-2498	310.6	1.72	30.9
GB009	-2522	310.2	1.94	30.6
GB010	-2547	310.3	1.71	31.9
GB011	-2502	310.4	1.89	30.8
Mean	-2533	310	1.8	31

TABLE 5.8: Average values over the trajectories of the potential energy (Epot), temperature (K), rmsd and radius of gyration (Gyration) for all the independent simulations of R136C at 310 K. For Emin the rmsd is calculated using the all-atom structure as the reference, while all the other values use the structure obtained from the energy minimisation (Emin) as the reference.

Replica	Epot [Kcal/mol]	T [K]	Rmsd [Å]	Gyration [Å]
Emin	-2641	-	4.46	31.3
GB000	-2547	310.2	1.86	31.1
GB001	-2537	310.3	1.81	31
GB002	-2552	310.4	1.83	31.2
GB003	-2580	309.8	1.82	31.4
GB004	-2553	310.1	1.86	31.2
GB005	-2592	309.9	1.88	31.2
GB006	-2538	310.3	1.91	31.1
GB007	-2537	310	1.7	31.4
GB008	-2539	310.8	1.81	30.9
GB009	-2571	310.3	1.79	31.6
GB010	-2574	309.7	1.85	31.3
GB011	-2567	310.1	1.75	30.5
Mean	-2557	310.1	1.8	31.2

5.4.4 R136C

R136C is a class II variant that was predicted to be damaging by SAAPpred with a confidence of 0.45. The cysteine (C136′) is replaced in the βD strand of the core of the N-terminal Rossmann-like domain and the all-atom simulations registered some distortions in the surrounding strands that can reach the co-enzyme binding site changing its geometry (Section 3.7). With lower potential energy, R136C is more stable, but the simulations took more time to converge compared with the wild-type (Figure 5.27).

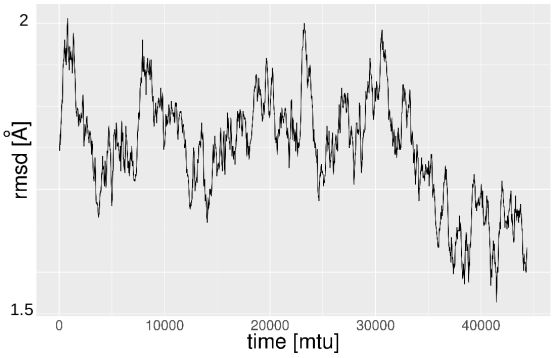
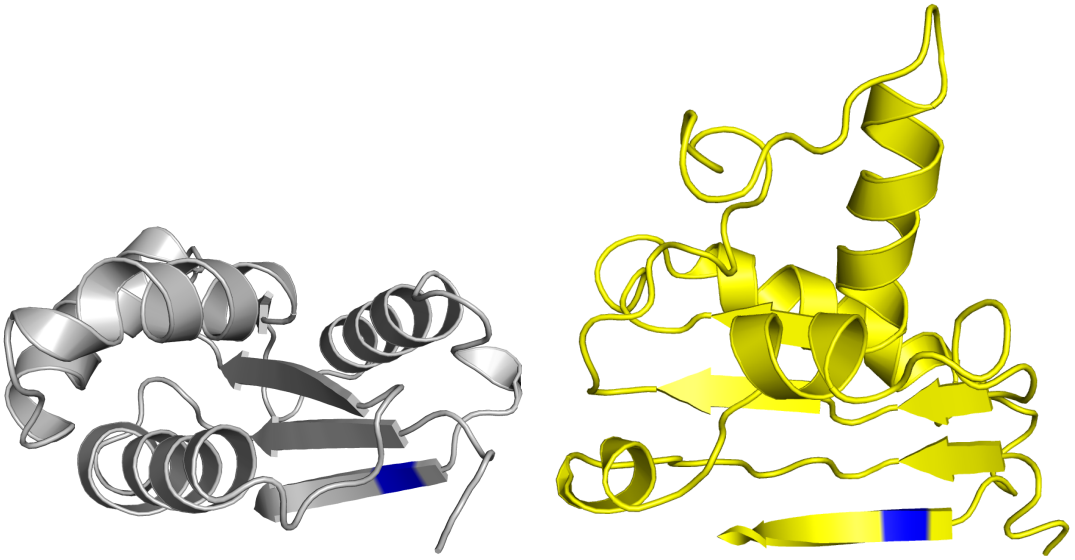


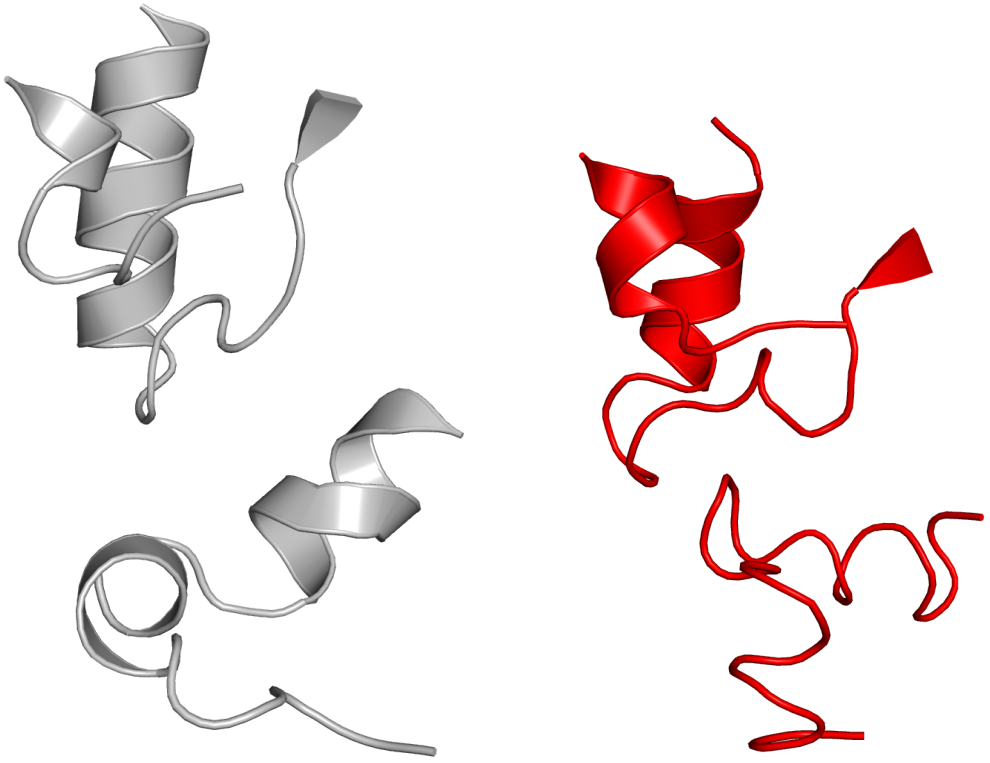
FIGURE 5.27: Rmsd profile of the replica 8 of R136C simulation, in which convergence happens only close to the end of the simulation.

The principal effect of C136′ is the disruption of the β sheet stability. The closer strand (βA) is completely distorted and shorter than the wild-type, while the other strand of the sheet (βB) breaks at the partner position to the cysteine (Figure 5.28b). It is possible that the hydrogen bonds of the arginine (R136) with the αc helix had an involvement in keeping the β sheet ordered, and that its removal destabilises the sheet stability.
In addition, the co-enzyme binding site, at the other extremity of the N-terminal Rossmann-like domain unfolded, but it is seen to be interacting with the unfolding helix spanning from G240 to G254 located outside the N-terminal Rossmann-like domain. This small helix unfolds and the resulting coil enters the co-enzyme binding site (Figure 5.29). The involvement of C136′ in this mechanism is not clear as it could either be that the unfolding of the N-terminus shifts the co-enzyme binding site towards the other end of the protein, or it could simply be the instability of the unfolding helix (G240-G254) which brings them close enough to interact.



(a) (b)

FIGURE 5.28: The N-terminal Rossmann-like domain in (a) the wild-type and (b) its distorted conformation in the R136C simulations. Residue 136 is in blue.

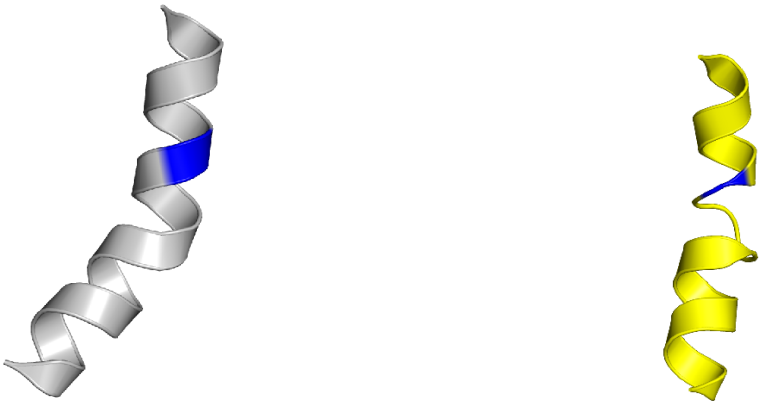


(a) (b)

FIGURE 5.29: The co-enzyme binding site region in (a) the wild-type and in (b) R136C. When the G240-G254 helix unfolds, it moves closer to the co-enzyme binding site, obstructing it.

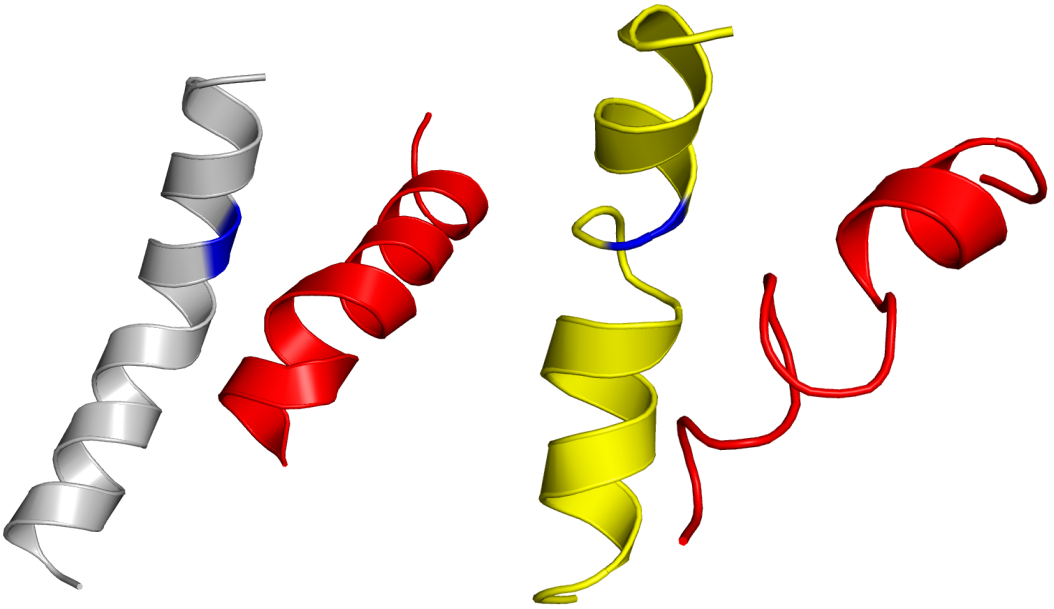
5.4.5 A461T

The A461T mutant is a known G6PD variant identified in the Yunnan chinese province whose reduced enzymatic activity has not been assessed yet, and it is therefore registered in the G6PD database of mutations [101] as non recorded (NR) . A461T was predicted as damaging by SAAPpred with a confidence of 0.23, and the all-atom simulations (Section 3.8) showed that T461′ causes a distortion in the adjacent αi helix thanks to interactions with the Q261 side chain. The αi helix constitutes the base of the G6P binding site, suggesting that T461′ could interfere with G6P binding in A461T. The UNRES simulations confirm this mechanisms and the following mechanism of action was proposed: the presence of the threonine (T461′) breaks the αn helix into two well-defined sections (Figure 5.30b), which in turn lead to the unravelling of the αi helix (Figure 5.31). The unravelling of αi increases the instability of the preceding G240-G254 helix, which eventually unfolds (Figure 5.32 in red). This helix moves more than the wild-type, at the point that it was observed, in one replica only, to enter and obstruct the co-enzyme binding site. A similar mechanism was observed in R136C, but here the unravelling of the helix and the effect of the mutation are directly connected. Overall A461T is less stable than the wild-type, the PES profile of the replicas shows the presence of several well-separated wells (Figure 5.33).



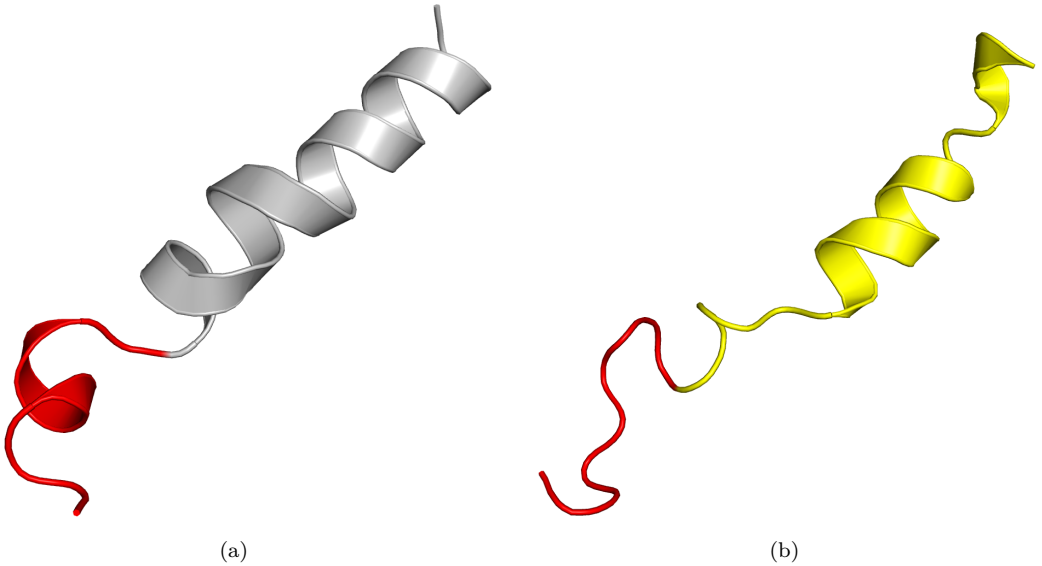
(a) (b)

FIGURE 5.30: The αn helix in (a) the wild-type and in (b) A461T. In the mutant, the threonine causes the helix to break.



(a) (b)

FIGURE 5.31: The αn and αi helices in (a) the wild-type and in (b) A461T. In A461T, the threonine (blue) in the αn helix (yellow) causes the αi helix (red) to unravel.



(a) (b)

FIGURE 5.32: The αi and the G240-G254 helices in (a) the wild-type and in (b) A461T. The unravelling of αi (yellow) speeds up the unfolding of the G240-G254 helix (red).

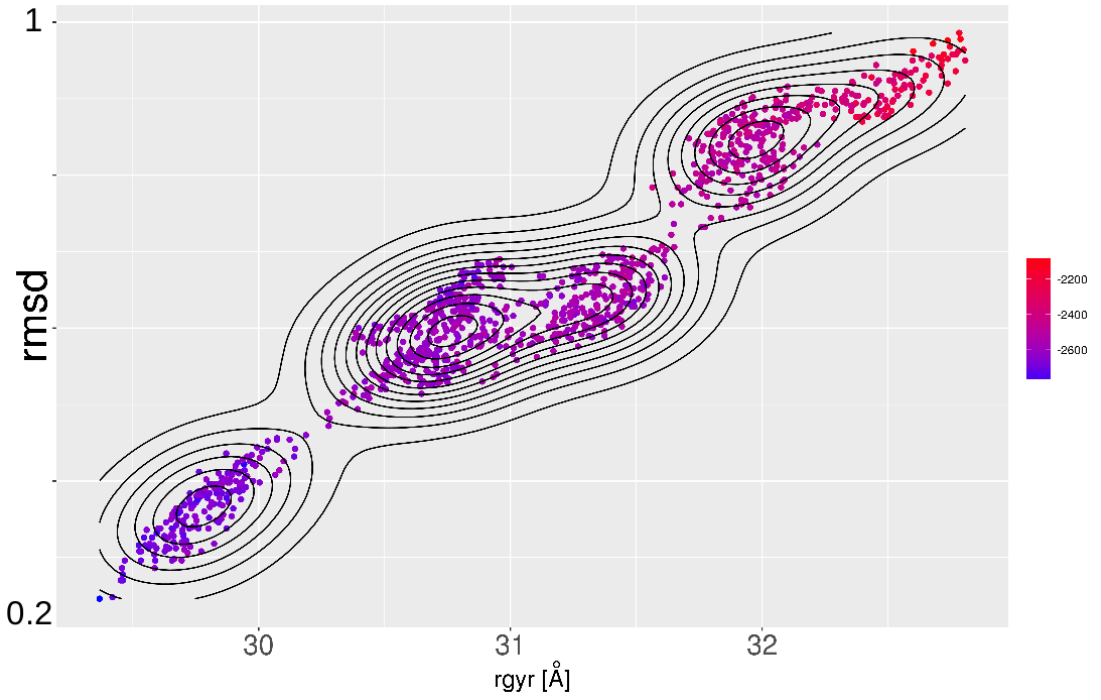


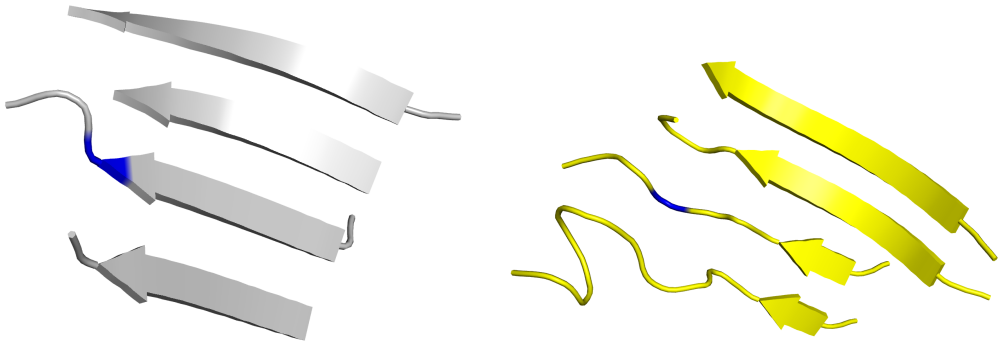
FIGURE 5.33: The PES sections explored by one of the replicas of A461T at 310 K. The radius of gyration is indicated with 'rgyr'.

TABLE 5.9: Average values over the trajectories of the potential energy (Epot), temperature (K), rmsd and radius of gyration (Gyration) for all the independent simulations of A461T at 310 K. For Emin the rmsd is calculated using the all-atom structure as the reference, while all the other values use the structure obtained from the energy minimisation (Emin) as the reference.

Replica	Epot [Kcal/mol]	T [K]	Rmsd [Å]	Gyration [Å]
Emin	-2611	-	4.46	31.3
GB000	-2514	310.1	1.75	31.4
GB001	-2553	310.1	1.82	31.1
GB002	-2559	310.4	1.88	31.2
GB003	-2521	310.2	1.6	31.5
GB004	-2522	310	1.8	30.5
GB005	-2498	310.3	1.2	31.6
GB006	-2528	310.2	1.9	31.6
GB007	-2527	310	1.8	31.2
GB008	-2553	310.2	1.77	30.8
GB009	-2532	310.1	1.9	31.5
GB010	-2566	310	1	31.5
GB011	-2520	310.2	1.6	31.4
Mean	-2532	310.1	1.7	31.2

5.4.6 UNRES simulations of mutants where no unfolding was detected in all-atom simulations

This section covers a set of mutants for which the all-atom simulations were not capable of detecting any specific damaging mechanisms. Mutant L140P, for example, has a proline inserted in the core of the N-terminal Rossmann-like domain, and was predicted to be damaging with a high confidence (0.78), but the all-atom simulations did not indicate any particular rearrangement or damage to the G6PD structure (Section 3.7). The UNRES simulations exposed the damaging mechanisms in a clear way. P140′ is located at the end of the βD strand of the N-terminal Rossmann-like domain, and the change in the backbone torsion angle caused by the proline produces a distortion in the closest strands of the sheet (βA and βE) which unfold (Figure 5.34).



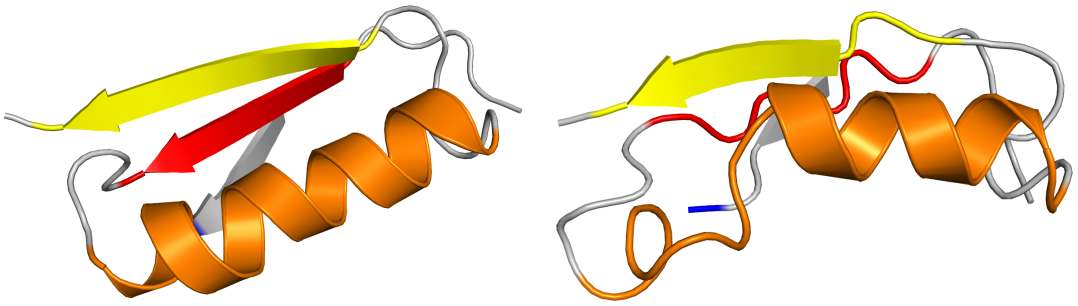
(a) (b)

FIGURE 5.34: The core of the N-terminal Rossmann-like domain in (a) the wild-type and in (b) P140′.

The unfolding is not complete and not dramatic, but it is capable of starting a series of modifications that can have more dramatic effects. The βA strand is directly connected to the αa helix (D42-R57), which sits below the two unstable helices of the domain (αb and αc). The unfolding of the βA strand, caused by the presence of P140′, affects the stability of the αa helix that in turn unravels (Figure 5.35).

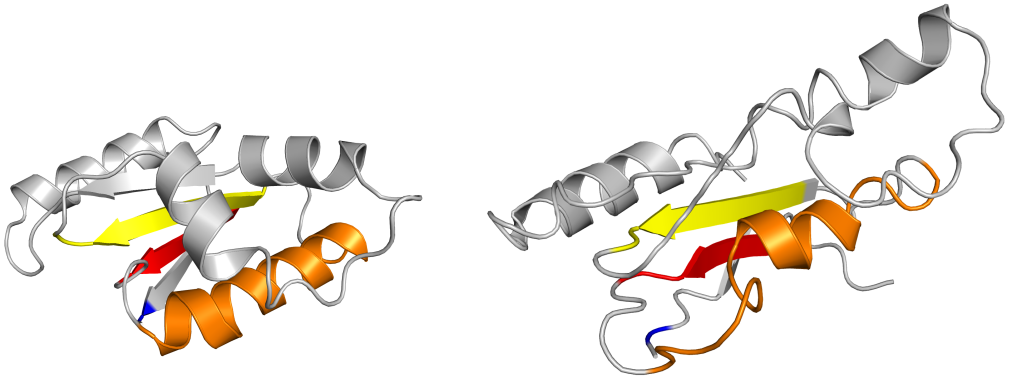
The most noticeable effect of this chain of events is the destabilisation of the top of the N-terminal Rossmann-like domain, which leads to a faster unfolding of the αb and αc helices (Figure 5.36). As a result, the substitution of L140 to P140′ is associated with local distortion that has the capability of propagating through the surrounding area, eventually accelerating the unfolding of the already unstable N-terminal domain. The fact that this mechanism was not observed in the all-atom simulations, gives confirmation of the limitation of the all-atom simulations in studying long-scale motion events.

The mutant A338E was predicted to be damaging by SAAPpred with a confidence of



(a) (b)

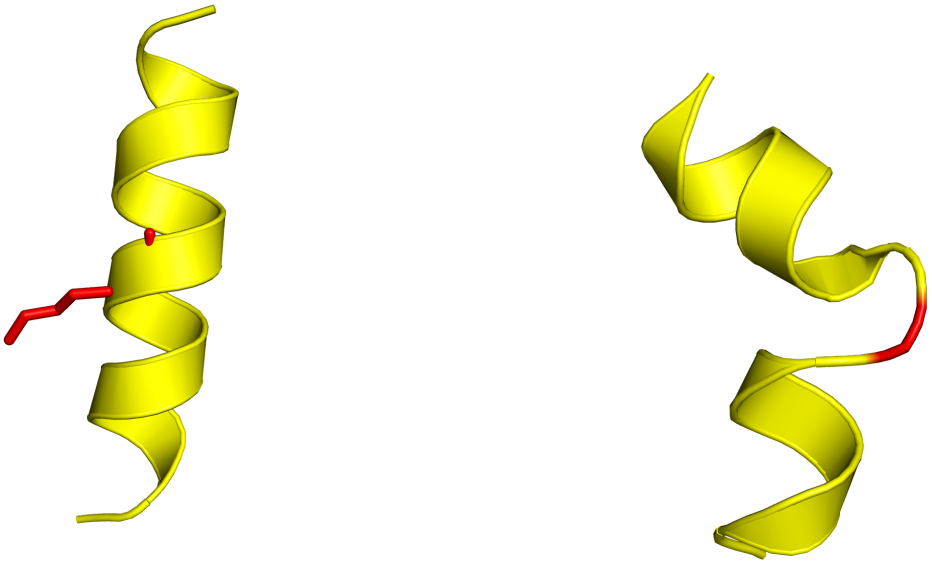
FIGURE 5.35: The two strands, βA (red) and βB (yellow), and the αa helix (orange) in (a) the wild-type and (b) L140P. Position 140 is in blue.



(a) (b)

FIGURE 5.36: The N-terminal Rossmann-like domain in (a) the wild-type and (b) in L140P. The unravelling of the αa helix (orange) increases the speed at which the top of the N-terminal Rossmann-like domain unfolds. P140′ is in blue and the strands βA and βB are coloured in yellow and red for comparison.

0.75, and it is located in the βI strand close to the C-terminus of the enzyme. The all-atom simulations detected an overall increase in instability of the C-terminal region, but did not demonstrate a clear mechanism of action of the mutation. The UNRES model shows how E338′ causes the unfolding of the βI strand increasing the motility of the C-termial region of the enzyme. The fact that E338′ sits in between the G6P and the structural $NADPH^+$ binding sites could suggest that motions in this region could affect both binding sites. Finally, E287K is another mutant predicted to be damaging (SAAPpred confidence of 0.68) which did not present any sign of different behaviour compared with the wild-type during the all-atom simulations (section 3.9). During the UNRES simulations, the αj helix, which contains K287′ breaks (Figure 5.37b) into two sections. In the most energetic replicas, this was enough to cause the partial unfolding of αj, while the other replicas did not unfold, but presented the damaged helix for the entire simulations.



(a) (b)

FIGURE 5.37: The αj helix as it appears at the end of (a) the all-atom and (b) an UNRES simulation. The helix breaks in two where K287′ (red) is.

5.5 Discussion

The decision to use the UNRES model was taken to overcome the sampling limitations of the classical all-atom model, which requires months of calculations for the collection of extensive simulation data. In contrast, the simplified UNRES model required a third of the time, and dramatically increased the sampling of the conformational space of G6PD. Almost all the simulations reached convergence and were capable of exploring several energy minima, making the study of unfolding dynamics at room temperature, for both G6PD and its mutants, possible. The UNRES model confirmed several of the findings of the all-atom simulations, especially the idea that the very local effects of a mutation can have an impact on the enzyme stability at a global level. The A^- mutant, for example, changes the structural equilibrium of the N-terminal Rossmann-like domain and induces its premature unfolding. The work with the UNRES model also made it possible to observe the damaging effects of those mutants for which the detection of any effect was difficult or impossible with the all-atom simulations (e.g. A338E or E287K). Unfortunately the loss of details of the simplified peptide chain representation increased the noise, making the detection of common unfolding behaviours among the mutants difficult, forcing a focus on a more global and general level. In the simulations, the effects of the mutations were clear, but it proved difficult to link those effects to specific structural movements. Nevertheless, these dynamics were important to confirm the damaging effects of the mutations and to prove that the overall damaging mechanisms proposed in Chapter 3 were indeed plausible.

Despite the satisfactory results obtained, at the end of the work done, it appeared clear that the UNRES model is not really suitable to detect the effects of single point mutations, but it is probably more suitable for other applications such as *de novo* folding or thermodynamics study of protein space, where it proved to perform well [150]. This consideration came after realising that the simplified model over accentuate the weakness of the G6PD structure, causing quick and different unfolding pathways. The result is the observation of the large scale movements that are likely to be caused by the mutations, without a clear and comprehensive description of the specific damaging mechanism at the origin of the global rearrangements observed. As a consequence of this last consideration, it was decided to focus on the collected all-atom data to have a better understanding of the damaging mechanisms of the G6PD variants. This work is discussed in Chapters 6.

Chapter 6

Additional studies: network analysis

The previous chapters illustrated how MD simulations were used to understand the effects of mutations on G6PD structure. All-atom simulations were capable of capturing the detailed local changes, but it was not always possible to produce as much data as would have been necessary. To try to overcome the sampling problem, it was decided to rely on a coarse-grained force field. The UNRES simulations allowed the comparison of global motions at a μs-time scale, but suffered from a great loss in detail. This lead to extensive data that were difficult to interpret. It was therefore decided to spend the last months of this project trying to extract, from the all-atom data accumulated, some clear explanations of the internal mechanisms that cause depressed activity in G6PD variants. For this reason, the main goal of the work presented in this chapter is to study how the damage propagates through the G6PD structure, based on the idea that structural effects can transmit through communication paths featuring correlated motions. The first section of this chapter describes how the G6PD structure was converted into a network of nodes and edges in an attempt to detect the important residues (hubs) in G6PD dynamics. In the second section, the G6PD network was used to find the shortest path lengths in both the wild-type and mutants, and to study how information (motion) propagation is altered in the mutants, compared with the wild-type. Ultimately the objective is to be able to find some examples that can demonstrate that, even if the damage is local, the changes induced in the mutants are capable of influencing binding sites behaviours. This could indicate that even if G6PD variants do not seem greatly to destroy G6PD structure, they can greatly affect its function.

6.1 Network analysis: Wordom

In the past decade, the interest in network-based analysis has grown steadily, and different methodologies have been proposed to study protein-protein interactions and metabolic networks. In more recent years, amino acid network (AAN [165]) models have been used to describe protein structures in an attempt to acquire a better understanding of proteins' topology [166], protein folding [167], protein signalling [168] and prediction of active centres and functions [169]. In all the different AAN formulations, Cartesian coordinates (PDB files) are used to reconstruct proteins as nodes (amino acids) connected by edges (amino acids interactions). The main difference between methods lies in how edges are defined. In the un-weighted edge models, distance-based functions define an edge only if the distance between nodes is less than a specific cut-off distance (R_c). Weighted-edge models, instead, combine network theory with dynamics information from other sources, such as MD experiments. Foor example *RING* is a web server that uses a combination of van der Waals contacts, atomic distances and alpha carbon distances to derive the amino acid network [170]. Alternatively, *RINerator* defines different interaction types based on the residues involved and builds an undirected weighted network with the weights of the edges that are proportional to the interaction strength [171]. Methods such as *JGromacs* [172], *NetworkView* [173] and *PSN-Ensemble* [174], are capable not only of building the networks from the PDB structure, but also of using MD simulation data to generate dynamics profiles of AANs. *xPyder* [175], for example, uses PyMOL to visualise common network parameters, such as hubs and intra/intermolecular interactions, focusing on improving the visualisation of AANs. All these AAN approaches have proven to be powerful tools in the study of protein behaviour and have found a wide set of applications. Topology AANs were used to determine the role of key residues in protein folding kinetics [176–178], to study the principles of protein-protein interactions [179–181], to identify functionally important sites [182, 183], and to predict thermal stability [180]. AAN analysis on G6PD and its variants was performed using the combination of both approaches implemented in *wordom* [184, 185]. Wordom uses an un-weighted edge model to build the Protein Structure Network (PSN) and a dynamics-weighted model, built combining PSN with cross-correlation information from molecular dynamics trajectories, to calculate the shortest path length between residues (PSNpath). Initially, all the frames of a trajectory are aligned to a reference structure (generally the first frame) and the average structure is calculated. Then, information on how the motions of the residues correlate with each other, is obtained using the Dynamic Cross-Correlation (DCC) algorithm [186], that assigns a score ranging from -1

(completely anti-correlated) to +1 (completely correlated motions) to each residue. At this point, the PSN of non-covalent connections between side chains is built as described in *Brinda et al.* [180]; each amino acid constitutes a node of the network that is connected to other nodes by edges based on the non-covalent interaction strength of their side chains. This interaction between two residues (I_{ij}) is calculated as

$$I_{ij} = \frac{n_{ij}}{\sqrt{N_i \times N_j}} \times 100 \tag{6.1}$$

where n_{ij} is the number of atom pairs between the side chains of residues (i and j) 4.5 Å apart, while N_i and N_j are the normalised factors for the residues type i and j [187]. Since the interaction strength I_{ij} depends on the property of both residues (i and j), the normalised factors takes into account the differences in the sizes of the side chains and their propensity to make contacts with other residues. The two residues are connected by an edge when their interaction strength (I_{ij}) is larger than an arbitrary cut-off value (I_{min}). At I_{min}, a transition occurs and a large number of weak interactions are lost. The correct value of I_{min} is set to the value at which the size of the largest cluster (number of nodes) is half the size of the largest cluster at I_{min} equal to 0. For G6PD, I_{min} was set to 3 and 2.9, depending on the temperature of the simulations; 310 K for the former and 400 K for the latter (Figure 6.1). Once the network is built, it is possible to calculate the shortest path lengths (L_{ij}) between nodes, defined as the size of the shortest edge that connects two nodes (i and j) [188]. Building the PSN may help the detection of those residues with high connectivity (hubs), that are therefore important for the protein activity or structure. These nodes provide robustness to the network against random mutations and help the linkage of different elements of secondary structure together [180]. By studying how these nodes are connected in the wild-type and in the mutants, it should be possible to observe the hidden changes in G6PD structure that may explain the relationship between structural local changes and the depressed activity seen in some variants. A schematic of the protocol is shown in Figure 6.2.

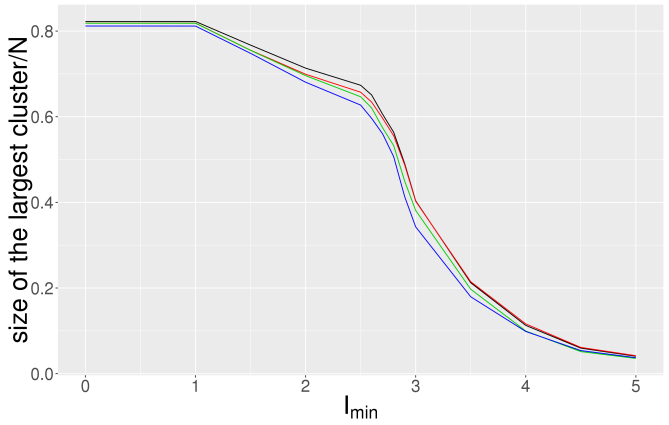


FIGURE 6.1: Size of the largest cluster as a function of I_{min} for the all-atom trajectories of the wild-type at 310 K (3 replicas in red, black and green) and 400K (1 replica in blue). The data were normalised by dividing the size of the clusters by the number of residues in G6PD (N).

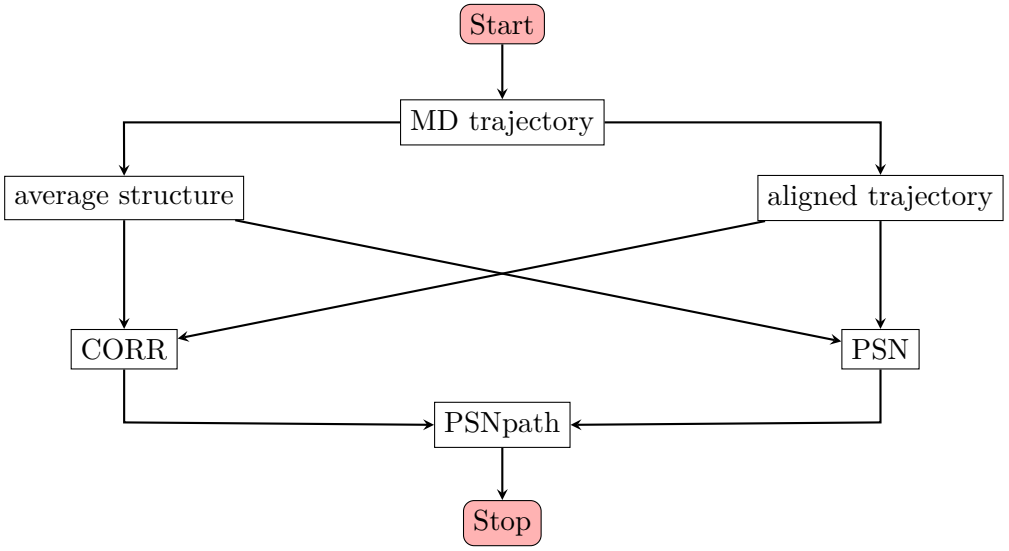


FIGURE 6.2: Diagram explaining the protocol used for the wordom analysis. Initially all the frames of the MD trajectory are aligned to the reference structure (first frame) and the average structure is extracted. This structure, together with the aligned trajectory, is used to generate the Protein Structure Network (PSN) and to calculate the correlation matrix residue-residue displacements (CORR) applying the Dynamic Cross-Correlation algorithm. These steps produces a series of data which include: the 3D representation of nodes and links, the average interaction strengths, the stable residue interactions, the hub frequencies and correlations, the stable cluster compositions and largest cluster size for each interaction strength (I_{ij}). Finally, on the basis of both PSN and CORR outputs, the shortest communication paths between hubs are generated (PSNpath). The described protocol is applied to all the trajectories individually, and the results are compared together. Because each trajectory was analysed individually it was possible to observe the changes in hub in different replicas of the same mutant, allowing a better comparison.

6.1.1 Protein Structure Network (PSN)

In G6PD, the total number of hubs (nodes with more than 4 connections with other nodes) existing in the wild-type was 45 at 310 K and 31 at 400 K (Figure 6.3), with 13 hubs corresponding of residues linked to existing G6PD variants. Of the latter, group only one residue is associated with a class III variant while the rest are class I or class II variants. 6 hubs are residues directly involved in binding (in either one of the binding sites) and 11 are adjacent in sequence to a residue directly involved in binding. More than half (6) of the residues involved in G6P binding are hubs, while for the co-enzyme and structural NADPH, the proportion of hubs is around one third (5 and 6 respectively). 3 of the detected hubs (Y70, R136 and R370) are at residues for which MD data were generated (Chapters 3 and 5). At a first look, these numbers seem to suggest that mutations in residues with high interconnectivity (hubs) are likely to be associated with a G6PD variant that exhibits a class I or II depressed phenotype, but the p-value calculated for the Fisher's exact test on the data does not provide any evidence against the assumption of independence of the classes, indicating that the hubs are likely to be evenly distributed between the classes (Table 6.1).

	Hubs	Non-Hubs		Fisher's exact test
class I/II	12	60	72	p-value= 0.1351
class III/IV	2	34	36	odds ratio= 3.368
	14	94	108	

TABLE 6.1: The Fisher's exact test was perfomed by comparing the number of hubs of the wild-type (Hubs) and the number of residues that are not hubs (Non-Hubs) for the different groups of classes. The test indicates that it is not possible to reject the NULL hypothesis which states that are not evenly distributed between the classes.

The PSN analysis was repeated for the mutants which were already considered in the previous chapters and their hub composition was compared to the wild-type using a multiple alignment approach (Appendix B). Overall, the total number of hubs and their location on the structure are maintained in the mutants, but it was possible to observe some local differences.

G306 is never a hub in the wild-type or in any mutants, but when the glycine is replaced by the arginine in G306R, not only does R306′ start acting as a hub (Figure 6.4), but also some changes are recorded at Y482 (Figure 6.5). These residues are far apart in sequence, but close in the 3D folded G6PD. MD simulations of G306R (Chapter 3 Section 3.7) have shown that R306′ (βH) interacts with I480 in the parallel

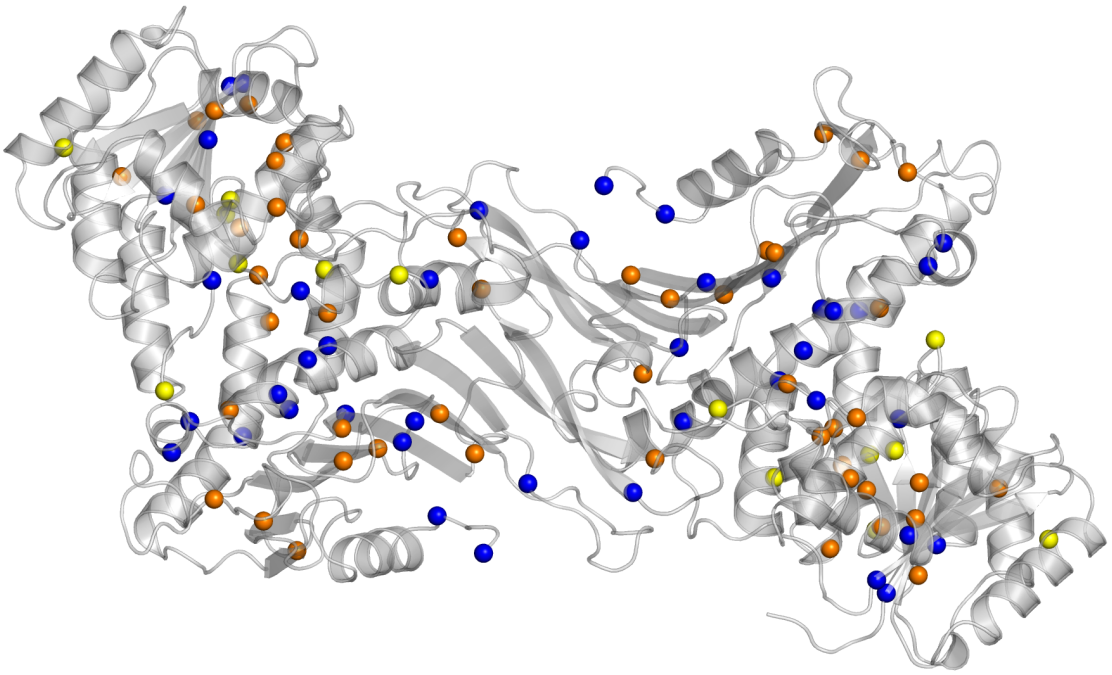


FIGURE 6.3: Total hubs detected in G6PD wild-type. In blue are represented the hubs only found in the trajectories at 310 K, while in yellow are coloured the ones for 400 K trajectories. In orange can be found the hubs shared between the two temperatures.

strand (βO). Even though I480 is not a hub, it is plausible that the R306-I480 interaction influences the adjacent residue (Y482) that in G306R stops acting as a hub in the region. The class II variant G306S, whose damage was difficult to explain in MD simulations, loses a hub at position 357, which is a residue that binds the structural $NADP^+$. The loss of a central-role residue, could be enough to lower the dimer stability.

```
wt      NSDDVRDEKVKVLKCISEVQANNVVLGQYVGNPDGEGEATK   320
G306R   NSDDVRDEKVKVLKCISEVQANNVVLRQYVGNPDGEGEATK   320
G306S   NSDDVRDEKVKVLKCISEVQANNVVLSQYVGNPDGEGEATK   320
        280         290         300         310         320

wt      HQIELEKPKPIPYIYGSRGPTEADELMKRVG   500
G306R   HQIELEKPKPIPYIYGSRGPTEADELMKRVG   500
G306S   HQIELEKPKPIPYIYGSRGPTEADELMKRVG   500
        470         480         490         500
```

FIGURE 6.4: Alignment of portions of the sequence (N280-K320 and H470-G500) of the wt and the mutants G306R and G306S, with the hubs highlighted in blue. (top) R306 is a hub only in G306R, suggesting that the arginine changes the balance of the are. (bottom) In G306R, the Isoleucine (I480) interacts with the new form hub (R306) changing the behaviour of the surrounding residues (H470 and Y482) which are no longer hubs.

G204 is located between two residues (Y202 and K205) that interact with G6P in the binding site. In the wild-type, the glycine does not behave as a hub and the mediator

of information (the closest hub) in the area is Y202 (H201 is only found as a hub in the wild-type and in R136C). The arginine in G204R has more atoms to share and becomes a hub, possibly changing the balance of the other hub in the area (Y202). A similar mechanism could explain the effects of the arginine in the G359R mutant. This residue is at the end of the βJ strand, in an area rich in residues involved in binding of G6P (K360, R365) or the structural NADP$^+$ (R357, N363, E364, K366 and R370). This area is surrounded by a set of four hubs; F354 and R357 from one side and R370 and Q372 from the other (Figure 6.6). When the small glycine is substituted by the large and polar arginine, the equilibrium of the region could change, modifying the capability of G6PD to relate with both G6P and the structural NADP$^+$. C358 and K360 act as hubs in some mutants (L140P, A338E, Y70H), further suggesting the importance of this region.

```
wt      RGSTTATFAAVVLYVENERWDGVPFILRC   358
A338E   RGSTTATFEAVVLYVENERWDGVPFILRC   358
       330          340          350
```

FIGURE 6.5: Alignment of portions of the sequence (R330-C358) of the wt and the mutant A338E, with the hubs highlighted in blue. In A338E, E338′ forces both F337 and C358 to behave like hubs, disrupting the hubs distribution found in the wild-type.

In A338E, for example, there is a swap of status of C358 and F337 that become hubs and R357 that instead loses that status. In addition, F337, the closest residue to the mutation, is close in 3D structure which binds G6P. The fact that F337 was not a hub in either the wild-type or in the mutants could be an indication that the changes could be a direct effect of the glutamate in A338E (Figure 6.6).

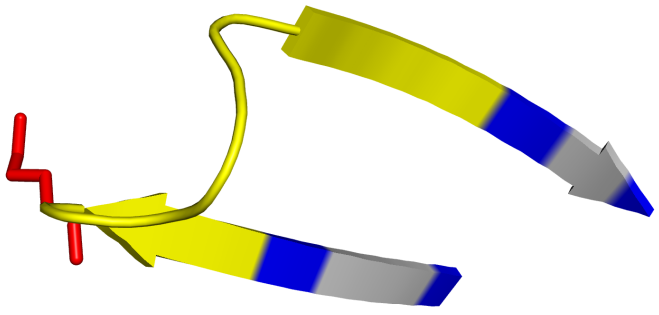
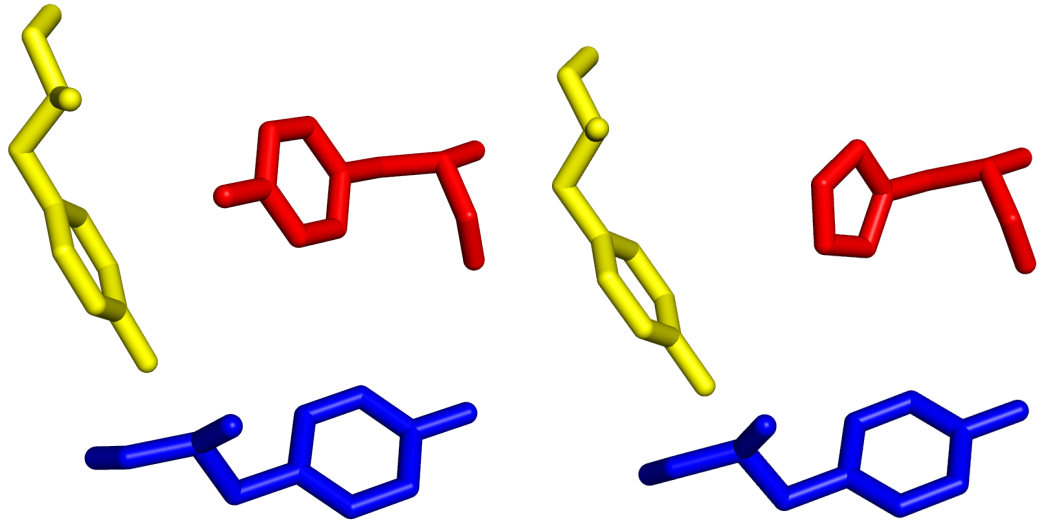


FIGURE 6.6: Representation of the area surrounding R359. In red, the hub K360 is indicated, and in yellow the area containing the binding residues is represented. The hubs (F354, R357, R370 and Q372) forming the extremities of the region are in blue.

Y70 is a conserved hub in the wild-type and in the mutants, but not in Y70H. Several residues close in sequence to H70′ (R72 and S73) bind the co-enzyme, suggesting that changes of status in the area could affect co-enzyme binding. This hypothesis is strengthened by the observation that while, in the wild-type, both Y70 and Y118 are hubs and

Y112 is not, in Y70H the roles are swapped and only Y112 is found as a hub (Figure 6.7). This is particularly relevant considering that Y112 lies inside the co-enzyme binding site and the observed changes could have consequences for co-enzyme binding. A similar



(a) (b)

FIGURE 6.7: Representation of the position of (a) Y70 and (b) H70′, in relation to Y112 (blue) and Y118 (yellow). When Y70 is mutated in H70′, Y112 acts as a hub and this change could affect the way Y112 binds the co-enzyme.

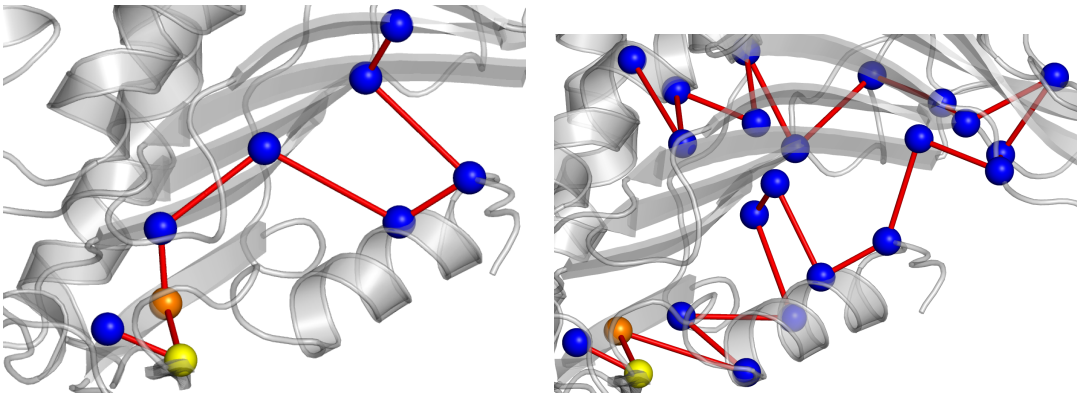
mechanism can be observed in R136C, where the changes in geometry observed in the MD simulations could be a consequence of interactions between R136, two hubs (F138 and Y139) and a set of three residues that directly bind the co-enzyme (A141-P143-P144-V146). As a final example, A461, which is never seen as a hub, forces Q261 to act as one when mutated into a threonine in A461T.

Some of the observations detailed above are a direct consequence of the way the network is built. In fact, because PSN uses the non-covalent interaction strength of the side chains to determine the importance of residues, the conservation or loss of a hub could be the mere result of the change in size (number of atoms) of the side-chains. Nevertheless, these results suggest that hubs are located in key regions of G6PD structure, and that the mutations have some role in how the hubs re-distribute in the surroundings. In several cases, these changes could be connected to the binding sites, possibly explaining the depressed activity.

6.1.2 Protein Structure Network Paths

The PSN data were further used to try to detect changes in the motion paths that are directly caused by the mutation. In proteins, some fast communication between sites is often required for crucial functions, but sometimes long paths are used to absorb or localize the effects of damaging perturbations. It is possible that a similar mechanism could help G6PD to dissipate the damage induced by a mutation by redirecting it along different paths. It also possible that non-damaging mutations (at the structural level) could reach important sites by blocking or altering specific paths. For the wild-type and some mutants, all the communication paths between the nodes found in the wild-type were generated. Because usually the shortest communication paths are likely to be the paths that more efficiently transmit information over long distances, the paths were then ranked based on length and interaction strength and only the shortest and strongest paths were considered. To observe only the likely alterations induced by the mutations, the paths between nodes close to the mutation site were analysed and compared with the data from the wild-type. In this analysis, examples were looked for that could prove the mutations' potential to affect G6PD function without destroying G6PD structure.

G306 is not involved in any paths in the wild-type, while in G306R, R306′ is associated with 6042 paths. The previous section discussed how Y482 does not maintain the status of a hub in G306R, and in fact in G306R there are half the number of paths connected to Y482. The effects of R306′ in terms of changing the motion dynamics in the area, can be understood by looking at a few examples. Y308 and R370 are hubs located close to R306′, both in sequence and in 3D structure. In the wild-type, R306′ and Y482 do not interact and Y308 and R370 are not connected, but in G306R they are, and their communication path (Figure 6.8a) goes through both Y482 and R306′.
A similar example is represented by H263, an important hub that binds G6P in the binding site. H263 and Y308 are not connected in the wild-type, but a 23 residue path connects them in G306R (Figure 6.8b). It seems that the ability of R306′ to interact with Y482 creates a triad of residues (Y308, Y482 and R306) that is capable of acting as a bridge between areas that are generally isolated in the wild-type. These new paths could interfere with the normal G6PD motions, eventually destabilising its structure. PSN path analysis was also used to try to find clues on differences in behaviour between G306R and G306S, two different mutations at the same residue. G306S is a class II variant that SAAPpred predicted damaging with a lower score than G306R (a non-existing variant). Even though the different nature of the two residues (serine and arginine) could be explanation enough, the MD simulations of G306S did not clearly demonstrate



(a) (b)

FIGURE 6.8: Examples of communication paths in G306R that do not exist in the wild-type: The path from (a) Y308 and R370 and (b) H263 and Y308. R306′ is in orange and Y482 is in yellow, while all the other residues involved are coloured in blue.

its damaging mechanism on G6PD structure (Chapter 3 Section 3.6). Similarly to what happened in G306R, G306S is also involved in communication paths, but fewer than G306R (6042 for G306R and 4606 for G306S). The comparison of the common paths shows that paths in G306R tend to be shorter than in G306S. For example to connect R357 (structural $NADP^+$ binding) to Y482, 9 residues are required in G306S and only 3 in G306R.

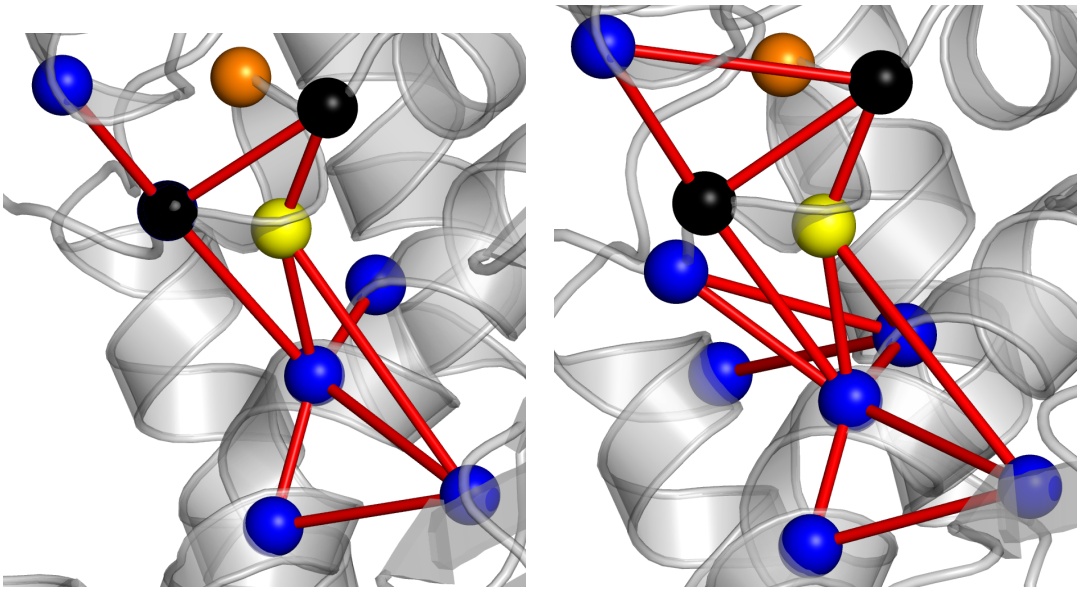
G306S ⇒ `R357-M496-I355-V341-L495-V304-I480-S306-A:Y482`

G306R ⇒ `A:R357-A:F337-A:R306-A:Y482`

Both G306R and G306S are capable of propagating motions in regions that are separate in the wild-type, but because the arginine side chain has more atoms to share, G306R is more efficient than G306S.

As described previously, G204 is a non-hub residue that is located in the G6P binding site, between two residues which bind G6P: Y202 and K205. In the wild-type, G204 is not a hub and there are no paths going through it, probably to avoid interference with movements of both Y202 or K205. When the glycine is mutated into an arginine, R204′ acts as a hub and plays a role in 3802 paths. To explore the possible effects that this may have, the paths passing through the close hub Y202 were checked and it was discovered that the mutation increases the connectivity of the network in the area close to the mutation site (Figure 6.9).

In the wild-type (Figure 6.9a), H201, Y202 and K205 are involved in fewer paths, contrary to G204R where the same residues interact more frequently with the surroundings. If a longer path is considered (W53-Y70), it is possible to notice how the paths diverge right after Y202 is reached:



(a) (b)

FIGURE 6.9: Increase in the number of the shortest communication paths for Y202 in the area around the mutation site for (a) the wild-type and (b) G204R. R204′ is in orange and Y202 is in yellow, while all the residues involved are coloured in blue. The two residues that bind G6P (K205 and H201) are in black.

wild-type ⇒ `A:W53-A:R57-A:W54-A:E438-A:D435-A:K205-A:Y202-A:F237-A:V259-A:I255-A:F250-A:D258-A:F253-A:P172-A:P144-A:Y147-A:L140-A:M37-A:Y70`

G204R ⇒ `A:W53-A:R57-A:W54-A:E438-A:D435-A:K205-A:Y202-A:H263-A:K171-A:D258-A:P172-A:F253-A:Y249-A:P143-A:V146-A:Y112-A:Y70`

In the wild-type, Y202 is oriented in such a way that its side chain points to the centre of the binding site and, in order to reach Y70, the motions first have to jump to F237 first and then to V259 (Figure 6.10 in red). In G204R instead, the Y202 side chain points to the bottom of the binding site, interacting with H263 (Figure 6.10 in yellow), resulting in a two residue shorter path (Figure 6.11).

This finding suggests that even if the arginine does not impact the structure at the enzyme on a global level, it forces a reorganisation of the surrounding residues that could impair the capability of key residues, such as Y202 and K205, to assume the correct position and orientation required for G6P binding.

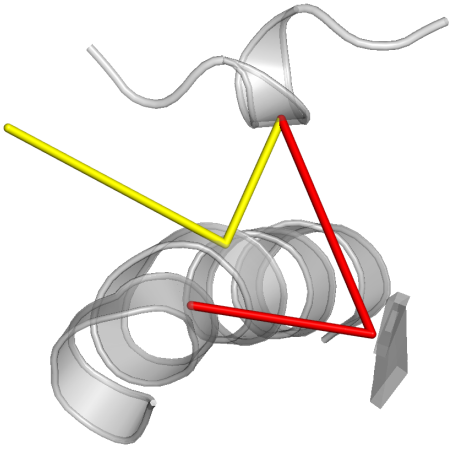
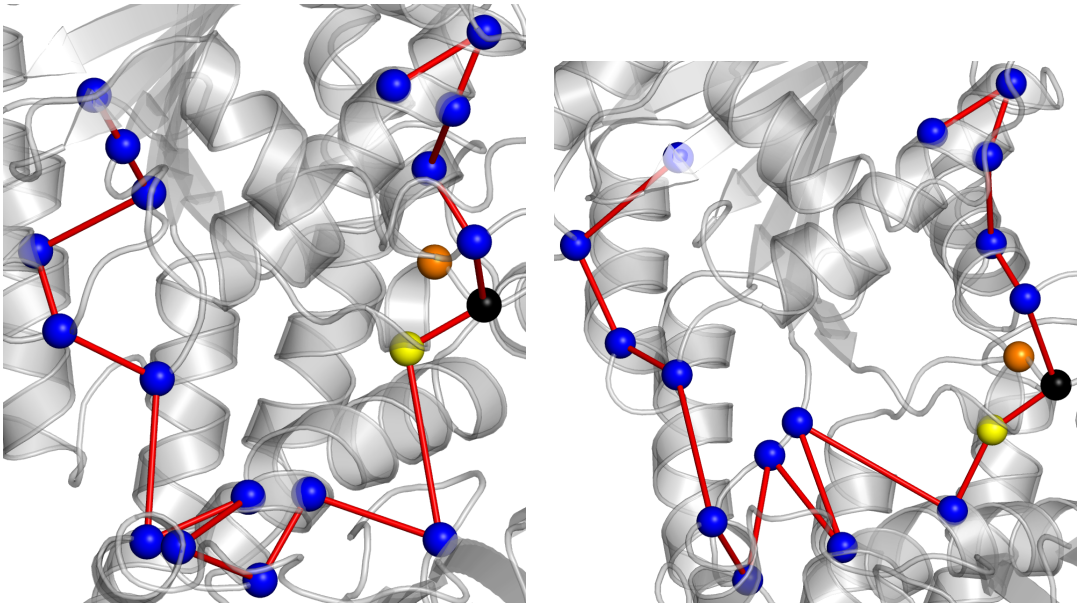


FIGURE 6.10: Schematics of the paths (W53-Y70) in the G6P binding site. In the wild-type (red line), the path goes through the binding site, while in G204R (yellow line), the biniding site is avoided.

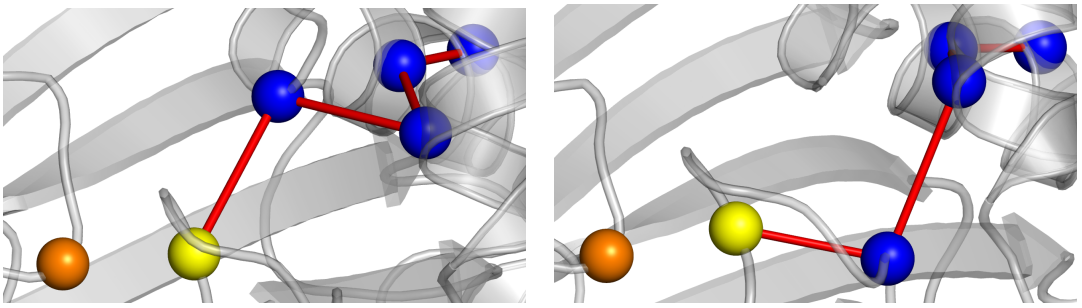


(a) (b)

FIGURE 6.11: Example of changes in path lengths between W53 and Y70 in (a) the wild-type and in (b) G204R. All the residues involved are coloured in blue, while R204 is in orange and Y202 is in yellow. K205, another hub that binds G6P, is in black.

Another example is represented by mutant R227Q. This mutant is located on the surface close to the binding site, it was predicted as damaging by SAAPpred with a low confidence (0.15), and is associated only with a mild depressed activity (class III). There were no significant differences in hub distribution compared with the wild-type, but while there are no paths passing through residue 227 in both the wild-type and R227Q, it was possible to detect some changes in the way W349, the closest hub in 3D structure, behaves. In fact, the mutant lost 200 paths and even though the maintained paths are

mainly identical in residue composition, they present differences in interaction strengths, suggesting that attractions and repulsions between Q227′ and W349 could cause W349 to move (Figure 6.12), altering the number of atoms available for the interactions (as fewer atoms means less interaction strength between close residues).



(a) (b)

FIGURE 6.12: Example of changes in path lengths between W349 and Q266 in the wild-type (a) and in R227Q (b). Q227′ is in orange and W349 is in yellow, while all the other residues involved are coloured in blue.

6.1.3 Discussion

A Network analysis was performed on both the wild-type and the mutants to explore how motions travel through the G6PD structure. Initially G6PD was rebuilt as a network of nodes and edges based on side-chain interactions and then these data were used to calculate the shortest communication paths between hubs. The network analysis was applied to G6PD and its mutants in an attempt of find evidence that the local structural changes of a mutation can propagate through the enzyme structure, affecting its functioning. All the observations obtained seem to confirm this behaviour. The mutations were found to be capable of altering the natural balances, by modifying the roles exerted by the residues in the area. The hubs are important residues of a protein, which has the function of linking together the different elements of its structure. Some mutations (G306R, G204R) introduce residues that direct the motions toward themselves, disrupting nearby hubs and potentially connecting isolated regions together. Other mutations (Y70H), weaken the mutated residue influence, isolating the area from the highly connected centres of the protein. All these changes are capable of reaching the binding sites, possibly altering the normal behaviour of these vital sites. The result is that in G6PD, the mutations are not capable of disrupting the global structure of the enzyme completely, but they tend to force local rearrangements and to generate small perturbations that influence the internal motion propagation, eventually causing the damaged phenotype recorded in the mutants.

Chapter 7

Summary and Conclusion

Malaria is a dangerous disease commonly transmitted by infected mosquitoes, and there is a strong connection between malaria and G6PD deficiency. In individuals with G6PD deficiency, the administration of the malaria drugs of the family of the 8-aminoquinolines may be very risky. To make malaria drugs safer, the development of low-cost and simple-to-use immunological field tests for G6PD-depressed-activity detection is required. The principal aim of this project was the study of the structure of G6PD and its variants to understand the structural effects of single point mutations on its stability. The collected information serves to assess the feasibility of using an immunological assay approach to detect G6PD variants, which are commonly associated with depressed activity.

Chapter 3 and Chapter 5 addressed the use of all-atom and coarse grained MD simulations in trying to identify common behaviours among G6PD mutants. Simulation at different temperatures and for different times made possible the detection of the local effects that are likely to be the origin of the depressed activity of the mutants. The collected data suggest that it seems to be unlikely for an antibody-assay to be developed to detect G6PD variants. It was hoped that G6PD variants would destabilise the protein leading to a partial unfolding at relatively high simulation temperature, and that the weak parts in the structure could be identified as common to several mutations. The effects of the mutations are in all cases very local and the mutations have a role in increasing the disorder in the surrounding area. This effect is obtained in different ways, but in almost all cases these effects can be connected to one of G6PD binding sites. In G306R, for example, the larger side chain of R306′ interacts with I480 causing a premature unfolding of the strands in an area close to the structural $NADP^+$ binding site. The threonine found in A461T is capable of deforming the αn helix at the base of

the G6P binding site, while in Y70H the presence of H70′ causes a distortion in size and geometry of the co-enzyme binding site. The only mutant that has a recognisable structural effect of the mutation is A^- . Here D126′ and M68′ destabilise the αc helix and the βB strand respectively, resulting in a specific unfolding pattern of the Rossmann-like domain of the enzyme. The coarse grained model (UNRES) made the unfolding, resulting from the destabilising interactions, visible, but the lack of details of the model made the identification of common behaviour among the mutants difficult. Even though the resulting effects of the mutants can be roughly expressed in similar manners (e.g. it destabilises the C-terminus of the enzyme), all the structural rearrangements observed in the mutants are specific to each mutation, and therefore the features expressed are not shared or maintained in other mutants. It is proposed that G6PD resilience to mutation damage, could be connected with the great importance that this enzyme plays in the cell. In fact, the complete deletion of G6PD in any cell of the organism is not compatible with life. To be functioning, G6PD must be stable and capable of reducing the damaging effects of possible mutations.

Particular care was used to study the behaviour of Pro172. Pro172 plays a key role in the correct positioning of both the substrate and the co-enzyme. In the literature, Pro172 was observed in both *cis* and *trans* form, so it was proposed that its function could be achieved by *cis-trans* isomerisation. To investigate if this is a real mechanism or the result of crystallography artefacts, the ω angle was monitored during the all-atom simulations. Even though Pro172 is in *cis* in all the simulations at low temperature (310 K and 400 K), the *trans* form of Pro172 was recorded at high temperature simulations (500 K), suggesting that the *cis-trans* isomerisation of Pro172 may occur in G6PD. Finally, further to understand the importance of Pro172 for the correct functioning of G6PD, it was decided to monitor the movements of Pro172 between the G6P and the co-enzyme binding site. In Chapter 4, metadynamics calculations were performed to study the evolution of the distances of the binding sites in relation to Pro172 (distances taken from the farthest of the residues that binds). Pro172 was found to oscillate between the two binding sites with a propensity to lean towards the co-enzyme binding site, suggesting that the movement of Pro172 moves the co-enzyme closer to the substrate binding site, allowing the interaction of K171 with G6P through its terminal amino group and with the co-enzyme through its carbonyl group.

Chapter 6 described some additional work that was done with the intention of better understanding the internal mechanisms of action of the mutations. In Chapter 5, network

analysis was used to have an idea of how the damage, caused by the mutations, propagates through the G6PD structure. Attention was given to highly connected nodes, that are important residues that help the connection between regions of the enzyme. Changes in these residues generally have a more severe effect because they play a central role in allowing communications (motions) to travel through the G6PD structure. By looking at the hub distribution in G6PD and its mutants, it was observed that mutations have the ability of altering the equilibria in the areas surrounding the mutation sites, resulting in differences in hubs' behaviours. The findings suggest that some mutants are capable of creating new paths and put in communication regions of G6PD that are isolated in the wild-type, while others induce a reorganisation of the information routes around the mutation sites. These perturbations impact the way information travels through the G6PD structure eventually leads to the phenotypic changes observed in G6PD variants.

Even though the development of a small number of antibodies capable of discriminating between mutants and wild-type upon binding was determined to be unlikely, this project outlined some key features that can help the understanding of this robust enzyme.

7.1 Future directions

This project started with the intention of studying as many G6PD mutants as possible, but because this number is close to 3500 possible mutants, and the fact that G6PD is a large system to simulate, only a small subset of 17 G6PD mutations was studied. The maim conclusion of this thesis is that the effects of the mutations are very local and, even though there are several similarities, it is not possible to find specific common behaviours between the mutants. Considering that a mass study is not feasible, it would be better to focus on the most severe (class I) or most common G6PD variants, in an attempt to find specific solutions for their detection. This could be possible for the A^- variant, which is one of the most common variants in Africa, or african ancestry, populations. For the other variants, a more detailed analysis could reveal some other mechanisms not identified in this project. All the simulations were performed in the absence of substrates; it would be interesting to see how differently G6PD and the mutants behave in the presence of both the G6PD and the co-enzyme. MD and free energy calculations may provide information on the affinity and binding capability in the mutants. This could prove to be challenging because of the absence of a complete parametrisation of the G6P molecule for the force fields considered, but would prove to be extremely helpful in assessing the changes in binding affinity of the mutants. Another possibility

would be the comparison between the monomer and the dimer. Only the dimer is recognised as being the active form of G6PD, but there are currently no explanations for this behaviour. Observing the dynamics of the monomer over time could help the understanding of the structural mechanisms that drive the dimer formation.

Appendix A

Scripting: doitGROMACS

A central role in the project was played by the shell script **doitGROMACS.sh**, written and maintained throughout the PhD. A normal GROMACS workflow sees the calling of several tools, each of them containing different parameters and files. The typing of these long commands is repetitive, slow and subject to errors. The script allows the automation of several steps, resulting in an increase in efficiency and precision, with a dramatic reduction in error rates. The first version of the script was very simple and contained only functions to equilibrate the system, run PCA and cluster analyses. Most of the parameters were hard-coded, all the prompts were passed using ***echo*** and there was no error control. Over time, the complexity increased to the point that the script required a complete rewrite before the final version (v 2.0) could be released. The factors that were considered important and which have driven the development of the code were:

- **A consistent naming system**: A non-trivial problem that is encountered while working with GROMACS tools is keeping track of all the simulation files. GROMACS generates several outputs with names which are set to default values (e.g. rmsd.xvg, ev1.pdb). A user can easily change them at will, but great care is needed to avoid confusion. This script addresses this problem by using some conventions in file names. The data are distributed in separate "temperature" directories, containing all the simulations at that specific temperature; inside, all the different simulation files are named using the following scheme: *NAMErX_TIME*; where NAME is the mutation (e.g.306r), X is the replica number and TIME is the simulation length in *ns*. Following this scheme, the file *400K/wtr2_800.xtc*, refers to the 800 ns long trajectory file of the second replica run for the wild-type at 400 K.

- **Universality**: Initially the script was tailored around the system and requirements of this project. Using the script on a different system would have meant manual editing of the script itself. This may be risky and is definitely not good practice, so, to allow other users to use the script, some changes had to be made. The problem was solved by using a configuration file to store most of the parameters and variables. A user will have to make changes only in this file leaving the rest of the script untouched. This approach not only allows other people to use *doitGROMACS.sh*, but it also improves readability and makes debugging easier.

- **Error handling**: The functionality of this code is guaranteed by its ability to pass files to and from different GROMACS tools. An error in one of these tools may generate a faulty file, causing another tool to fail. Error handling functions stop the execution as soon as an error is detected. Furthermore, to help with the debugging process, specific *.log* and *.err* files are generated for each executed function.

A look at the evolution of the script over time, reflects how its changes are not simply an addition/removal of functions, but a dynamic process that ran parallel to the evolution of my personal programming skills. Slowly, the script evolved from a list of repetitive ***echo*** statements and ***if*** blocks, to a well structured and organised collection of optimised functions.

The script is mainly written in *bash*, with additional functions both in *R* and *perl*. This script is accessible from github at the following address:
https://github.com/frac2738/doitGROMACS_final.

A.1 Main script (doitGROMACS.sh)

The main script has the role of linking everything together by taking the user commands and passing them to the required functions. The script reads most of the parameters from two distinctive sources: flags and a configuration file. The main difference between the two is that flags are used to define *function independent* variables, while the configuration file contains parameters that are specific and used only by certain functions (*function dependent* variables). The flags are set from the console when the script is invoked and define names, executable paths and files. The complete list of flags available can be obtained with the command `doitGROMACS.sh -h` :

```
Option    Type      Value       Description
------------------------------------------------------------------------
-[no]h    bool      yes         Print help info
-g        bool      bool        Set gromacs 5 syntax

                                        [ ALWAYS REQUIRED ]
------------------------------------------------------------------------
-b        string    acrm        Set the location of binaries
                                acrm      - Darwin building computer
                                emerald   - Emerald cluster
                                default   - standard linux computer
                                The user can add more machines in the
                                binary configuration file.
-n        int       wt          Set the name

                                        [OPTIONAL (function dependant)]
------------------------------------------------------------------------
-t        int       200         Set the simulation length
-s        string    .tpr        .tpr file
-f        string    .xtc        Trajectory file
-c        string    .pdb        Pdb file to use to start a simulation
-e        string    .edr        Energy file
```

The flags -b and -n are always required and their existence is checked when the script is invoked. All the other flags are function dependent and they are checked only when a specific function is called. When a function is called without the correct flags, the execution is halted and a warning message indicates which flag should be set.

`16:34:45: execution halted: tpr and trj files are required (set them with -s -f)`

The configuration file (doitGROMACS.config) contains all the variables that are required by the different GROMACS tools. The force field used, the shape of the simulation box and the simulation temperature are some of the existing options.

```
...
optionFF='6'           #   force field: AMBER99S-ILDN protein, nucleic AMBER94
optionWM='1'           #   water model: TIP3P
optionBOX='triclinic' #   box shape
optionDISTEDGE='1.4'   #   protein-box distance [nm]
optionTEMP='400'       #   system temperature [K]
...
```

To allow the use on different machines, the configuration file also contains the location of the different GROMACS tools and R binaries. When *doitGROMACS* is invoked, the value of the -b flag is read and doitGROMACS checks that there is consistency between this value and the binaries written in the configuration file. If the configuration file points to the wrong binary, doitGROMACS will use standard linux binaries for both GROMACS and R. At the moment 3 machine classes are supported: any linux machine in the UCL network (*acrm*), the UCL cluster Emerald (*emerald*) and a general linux machine (*standard*). The user can easily defines its own machines, by adding them to the configuration file. For example, to use a different version of R, the user can add `RscriptEXE_hello='/export/user/R-v1.3/bin/Rscript'` to the configuration file and then use `-b hello`.

Consistency received a lot of attention and several mechanisms assure that each function is executed only when the required files and flags have been specified. Once everything is checked, the script prints the available options and waits for the user input.

```
                    ----------------------------
------------------- GROMACS 4.6 syntax will be used --------------------
                    ----------------------------
                             -----------
--------------------------- doitOPTIONS ---------------------------
                             -----------

    all           - Starting from scratch **
    emin          - Starting from E-minimization **
    nvt           - Starting from NVT **
    npt           - Starting from NPT **
    h20           - Remove water from a trajectory file
    cond          - Check the simulation conditions (U-T-P-density)
    rmsdfg        - Calculate RMSD, GYRATION RADIUS and RMSF [backbone & sidechains]
    dssp          - DSSP analysis
    dssp_perc     - percentage of ss using a xpm file
    contact       - Compute contact maps
    cluster       - Cluster analysis
    pca           - PCA analysis
    sas           - SAS analysis
```

```
        sas-sites     - SAS analysis on only the binding sites
        hb            - Hydrogen bonds analysis [not yet implemented]
        hb-sites      - Hydrogen bonds analysis on binding sites
        ggplot        - Plot with ggplot (R)
        indexCreator  - Create binding sites index for the mutant
        omega         - Calculate omega values for all the residues

        mean          - calculate the mean of values [2nd column]
        mean_multi    - calculate the mean of values [several columns]
        modvim+       - replace "@" with "#" in a file
        catomain      - rebuild a full atoms structure from CA structure
        split_states  - given an unres trj extract all the frames and convert
                        into all-atom structures

    ------------------------------------------------------------------
                               -----------

    **  Options that require a parameter file (.mdp) that MUST be placed in the
        same directory; the functions only accept these files:

                            TEMP-min.mdp
                            TEMP-nvt.mdp
                            TEMP-npt.mdp
                            TEMP-md.mdp

        with TEMP = temperature in Kelvin (310, 400, ...)
        The doitGROMACS tarball contains some examples of mdp file.

execute option :
```

Examples on how to run the script can be found in section A.4.

A.2 Functions

A.2.1 Equilibration

- inputs(): function that prepares the simulation box;
- energy_minimization(): function that minimises a protein;
- nvt(): function that equilibrates in an nvt ensemble;
- npt(): function that equilibrates in an npt ensemble;

This file groups the functions designed to prepare the system to a production MD simulation. The function *inputs()* takes a PDB file (-c flag) and creates a topology for that protein. It then builds and solvates a box around the protein and, finally, adds ions to

neutralise the total charge of the system. The energy minimisation, and the temperature and pressure equilibration steps are held by the functions *energy_minimization()*, *nvt()* and *npt()* respectively. These last functions simply create a starting parameter file (.tpr) from a .mdp file (specified in the config file) and run mdrun from the GROMACS suite. For an example of how these functions work see section A.4.1.

A.2.2 Analyses

A big part of doitGROMACS.sh is represented by the collection of functions designed to run the analyses that are generally used to analyse a trajectory (.xtc). At the moment, the script supports the following functions:

- h20: removes the water from a trj file;
- cond: prints the simulation conditions (Potential,Temperature,Pressure, Density);
- rmsdf: calculates rmsd, rmsf and radius of gyration;
- dssp: assigns secondary structure to each frame of the trajectory;
- cluster: runs cluster analysis;
- pca: runs PCA analysis (C-alpha);
- sas: runs SAS analysis;
- hb: counts hydrogen bonds;
- ggplot: plots the outputs of the other functions with ggplot;
- indexCreator: generates index files for a subset of atoms specified in the configuration file.
- catomain: rebuilds a full structure from a Cα-only structure.
- rama: calculates ψ, ϕ and ω values for the residues.

A.2.2.1 h20

This function removes the water molecules from a trajectory and corrects any artifacts generated by the periodic boundary conditions (pbc). The function first removes all the water coordinates from the reduced precision trajectory (.xtc), and it then does the same for the input file (.tpr) and creates a structure file (.gro) with the first structure in the trajectory. This structure can be loaded into programs such as vmd or pyMOL for visualisation. Finally, it re-centres the protein in the box and accounts for the periodicity by removing the translations and rotations.

A.2.2.2 cond and rmsdf

These functions have the main objective of extracting simulation conditions and energy from a trajectory. The function *cond()* extracts the temperature, potential, pressure and density profiles from the energy file (.edr). By observing how these values change during the simulation, it is possible to predict and avoid large changes in the forces that lead to system instabilities and simulation failures. The *rmsdf()* function calculates the rmsd, radius of gyration and rmsf for the selected trajectory. For an example of how these functions work see section A.4.2.

A.2.2.3 dssp

This function uses the *dssp* program [189] to assign the correct secondary structure to each frame of a given trajectory.

A.2.2.4 pca

This function can be used to perform PCA analysis by calculating and diagonalising the mass-weighted covariance matrix. Additionally, *pca*() uses the projections of the first two eigenvectors to make a multi-dimensional free-energy plot. The raw plot is then enhanced in R, using the *ggplot()* function (A.2.2.8). For an example of how these functions work see section A.4.3.

A.2.2.5 cluster

This function executes cluster analysis using a clustering algorithm specified in the configuration file (default is set to *gromos*). To avoid overcrowding of the main directory, the function first creates a new directory, checks if an rmsd matrix already exists and then runs the cluster algorithm. This simplifies re-running the analysis with a different algorithm, because no time is wasted recalculating the matrix if it is already there.

A.2.2.6 sas and hb

These functions are used to calculate Solvent Accessible Surface area and to count hydrogen bonds. There are two version of these functions: one performs the analysis on the entire protein (sas and hb), while the second only on specific groups defined in the

configuration file (sas-sites and hb-sites). In this project, these second functions were used to observe the changes in the binding sites of G6PD. All the parameters, such as the size of the probe or the number of dots per probe, are defined in the configuration file.

A.2.2.7 rama

This function calculates the values of the ϕ, ψ and ω dihedral angles for each residue in a trajectory. The function calculates both averages over the trajectory and the values in each frame. Additionally, it allows the user to specify a residue (e.g. pro-172), and extract and plot the Ramachandran plot and ω distribution for that residue only.

A.2.2.8 doitRGROMACS

doitRGROMACS is an R script which contains all the subroutines used to plot (using *ggplot*) the output files from the analyses. It is generally called at the end of one function and it checks what type of file has been passed (rmsd or rmsf) and it calls the function required by that specific file type.

A.3 Error handling

DoitGROMACS.sh works by executing several commands in sequence, meaning that, if pipelines are not stopped when errors arise, corrupted files are passed to the next tool. *doitGROMACS.sh* has several functions that check and guarantee that everything runs smoothly. One of the mechanisms used is the checking of the exit code of commands. The script catches the exit codes ($?) and halts the execution whenever the values differ from *zero*. A prompt tells the user what has happened and details are written into the *.err* file. thus making debugging easier.

```
checkExitCode() {
  exitvalue=$?
  if [ ! $exitvalue -eq 0 ]; then
    error_exit " The last function returned the wrong exit code, execution halted.
                 Check logs for further details."
  fi
}
```

A.4 Examples

A.4.1 System equilibration

System equilibration is generally the first thing that is done before starting a production dynamics run. In these initial steps, the simulation box is created, solvated and all the unfavourable geometries and interactions inside the box are removed. Temperature and pressure are then applied until the system is stable. All this can be achieved using doitGROMACS.

```
doitGROMACS.sh -b standard -n 2bhl -c 2bhl.pdb -g
```

The -b flags tells doitGROMACS which machine the user is using (a standard linux machine in this case), while the -g flag indicates that the user wants to use the newer version of GROMACS (5.0.4 or above). The value of -n sets the PREFIX used for all the output files (e.g. 2bhl_processed.pdb or 2bhl_min.xtc). When the *all* option is selected at the prompt, doitGROMACS starts building the simulation box around the protein in the PDB file specified with the -c flag. After creating the topology, building the box and solvating the system, doitGROMACS prompts the user for the number of ions required to neutralise the total charge of the system.

```
NOTE 3 [file topology.top, line 46]:
  System has non-zero total charge: -7.999994
Specify the number of ions to be add added to the system [+/- integer]
(e.g. + 12 or - 23) + 8
```

A message lists the changes and doitGROMACS starts minimising the system.

```
Back Off! I just backed up topology.top to ./#topology.top.2#
Replacing solvent molecule 5695 (atom 32505) with NA
Replacing solvent molecule 15786 (atom 62778) with NA
Replacing solvent molecule 12443 (atom 52749) with NA
Replacing solvent molecule 51683 (atom 170469) with NA
Replacing solvent molecule 52802 (atom 173826) with NA
Replacing solvent molecule 9001 (atom 42423) with NA
Replacing solvent molecule 18792 (atom 71796) with NA
Replacing solvent molecule 2490 (atom 22890) with NA
```

When the energy has converged, doitGROMACS calls, in sequence, the functions *nvt* and *npt* that equilibrate the now minimised system first in an NVT and then in an NPT ensemble. At the end of every step, doitGROMACS outputs some files (2bhl_potential.xvg, 2bhl_density.xvg, 2bhl_pressure.xvg and 2bhl_temperature.xvg) that help the user checking the stability of the system along the equilibration process (Figure A.1).

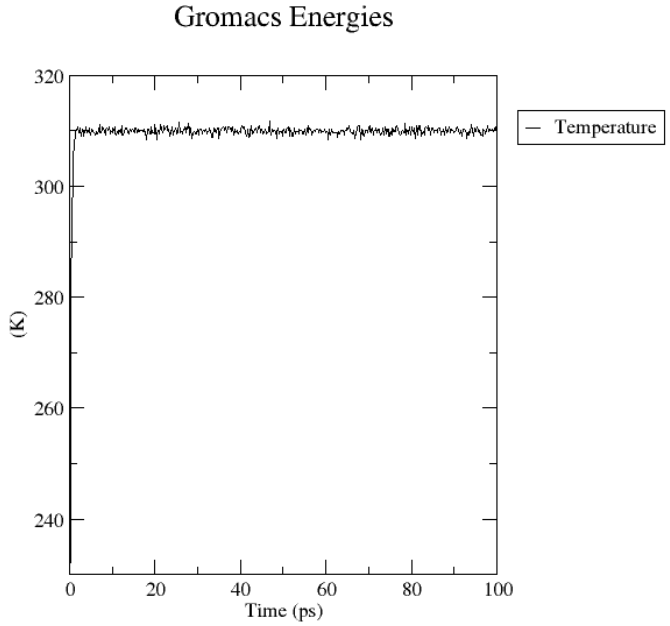


FIGURE A.1: Default representation (using *xmgrace*) of the temperature profile of the nvt equilibration step for the wild-type.

If the system is not well equilibrated, it is possible to rerun some of the previous steps until convergence is reached. This can be achieved by invoking:

```
doitGROMACS.sh -b standard -n 2bhl -c 2bhl.pdb -g
```

and selecting an option between *emin*, *nvt* or *npt*, depending on the step which needs to be rerun. If *emin* and *nvt* are selected, doitGROMACS will take care of completing the equilibration of the system until the end. The force field, the water model and all of the other parameters are defined in the 'simulation options' section of the configuration file.

```
#================================= simulation options ========================

optionFF='6'           #  force field: AMBER99S-ILDN protein, nucleic AMBER94
optionWM='1'           #  water model: TIP3P
optionBOX='triclinic'  #  box shape
optionDISTEDGE='1.4'   #  protein-box distance [nm]
```

Additionally, some GROMACS tools require a specific parameter file (.mdp) for their functioning. These files must be specified in the 'MDPs option' section of the configuration file.

```
#================================== MDPs options ==============================
optionTEMP='310'      #  system temperature [K]
ioniMDP='./ionization.mdp'
minMDP="./"$optionTEMP"-min.mdp"
nvtMDP="./"$optionTEMP"-nvt.mdp"
nptMDP="./"$optionTEMP"-npt.mdp"
```

A.4.2 rmsdfg

This section describes how to use doitGROMACS to extract the rmsd, rmsf and radius of gyration profiles from a trajectory and plot them using ggplot. This can be done by invoking doitGROMACS and selecting the option *rmsf* at the prompt:

```
doitGROMACS.sh -b acrm -n 2bhl -t 150 -e 2bhl_150.edr -s 2bhl_150.tpr
-f 2bhl_150.xtc -g
```

During MD experiments, GROMACS records information about all the energies of the simulation (potential energy, temperature, pressure, ...) in an .edr file. This file (2bhl_150.edr in the example) is passed to doitGROMACS using the -e flag. To work properly, doitGROMACS needs two additional files: 2bhl_150.tpr and 2bhl_150.xtc. The first contains the simulation parameters (starting structure and molecular topology) and the latter stores the coordinates of the system during the entire trajectory. After checking the existence of the required files, rmsdfg calculates in sequence rmsd, radius of gyration and rmsf. The user has the option of deciding which group to use for the calculation by editing the configuration file.

```
optionRMSD='Backbone'
optionGYRATION='Protein'
optionRMSFb='Backbone'
optionRMSFsc='SideChain'
```

At the end of the execution, doitGROMACS outputs 4 files named after the values of the flags -n and -t (2bhl_150 in the example). These files are:

- 2bhl_150_rmsd.xvg: rmsd profile;
- 2bhl_150_rgyr.xvg: radius of gyration profile;
- 2bhl_150_rmsf_bb.xvg: rmsf profile of the backbone;
- 2bhl_150_rmsf_sc.xvg: rmsf profile of the side-chains.

These files can be visualised with the software *grace* (xmgrace) (Figure A.2) or they can be further processed by doitGROMACS to obtained enhanced figures (Figure A.3).

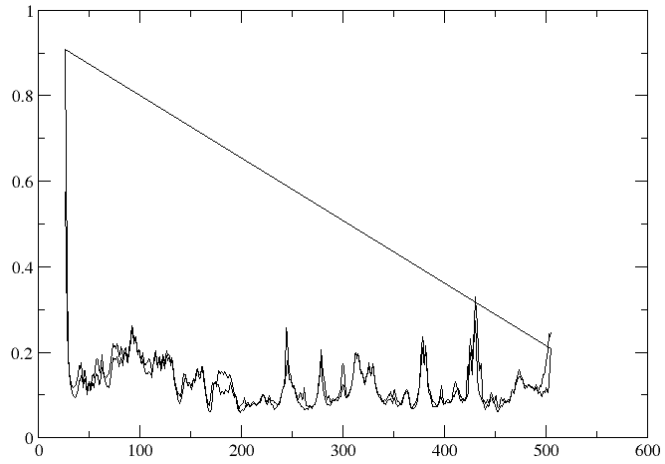


FIGURE A.2: This figure presents an example of an rmsf profile visualised in *grace*. In the default representation, the two chains are connected together resulting in a diagonal line on screen.

This can be achieved through choosing the *ggplot* option at the prompt. This function calls an R script (doitRGROMACS.R) that produces the required figure using ggplot .

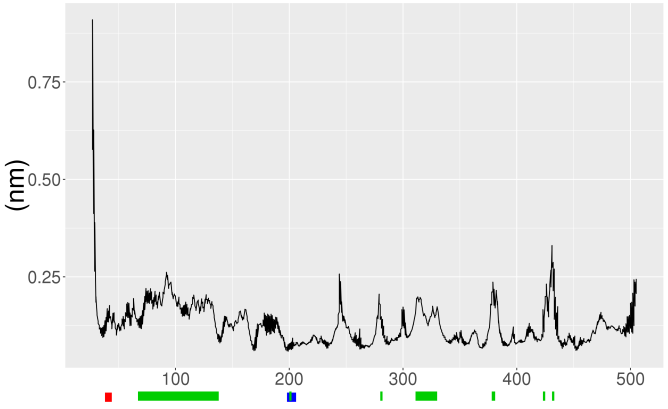


FIGURE A.3: This figure presents the same rsmf profile as Figure A.2, obtained by running doitRGROMACS.R.

A.4.3 PCA

The doitGROMACS function dedicated to PCA is called *gromPCA()* and can be used by invoking doitGROMACS and selecting the option *pca* at prompt:

```
doitGROMACS.sh -b acrm -n 2bhl -t 150 -e 2bhl_150.edr -s 2bhl_150.tpr
-f 2bhl_150.xtc
```

gromPCA first looks for a directory named PCA_2bhl, creates one if it is not there, and then checks the existence of the covariance matrix (covariance.xpm) which captures the correlation of motion between atom pairs. This step is necessary to avoid having to recalculate the matrix every time the same trajectory is analysed. If the matrix is not found, *gromPCA* calculates it using the C_α atoms if not otherwise specified in the configuration file:

```
optionSTARTime='0'        #  [ps]   [option used for cluster/pca/dssp/sas]
#================================== pca options ============================
optionPCA='C-alpha'  #
optionDTpca='100'    #  jump in [ps]
```

doitGROMACS allows the user to skip some frames of the trajectory and to run the analysis only on a portion of the trajectory only. This is done by changing *optionSTARTime* and *optionDTpca* in the configuration file. The covariance matrix now calculated is diagonalised to obtain eigenvectors (collective motion per atom) and eigenvalues (value of the involvement of the atom in the motion). Generally, only the first eigenvectors capture the important motions of the protein, this is why the user is required to input the number of eigenvalues to study.

`how many eigenvalues do you want to use for the analysis?`

The choice is facilitated by a pop-up window which presents all the eigenvectors and their respective amplitude (Figure A.4).

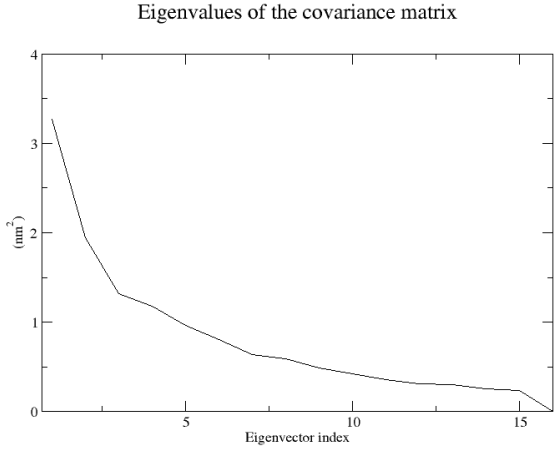


FIGURE A.4: Default visualisation of the eigenvalues plot with grace.

For each eigenvalue selected, doitGROMACS calculates its projection over time and it extracts the extreme structures from the trajectory, saving them in a pdb file (ev1.pdb, ev2.pdb, ...). These PDB files allow the observation of the motion along the cartesian axis in PyMOL or any other visualisation program (Figure A.5). Finally, *gromPCA*

plots the projections of the first two eigenvectors against each other to obtain the PES profile (Figure A.6).

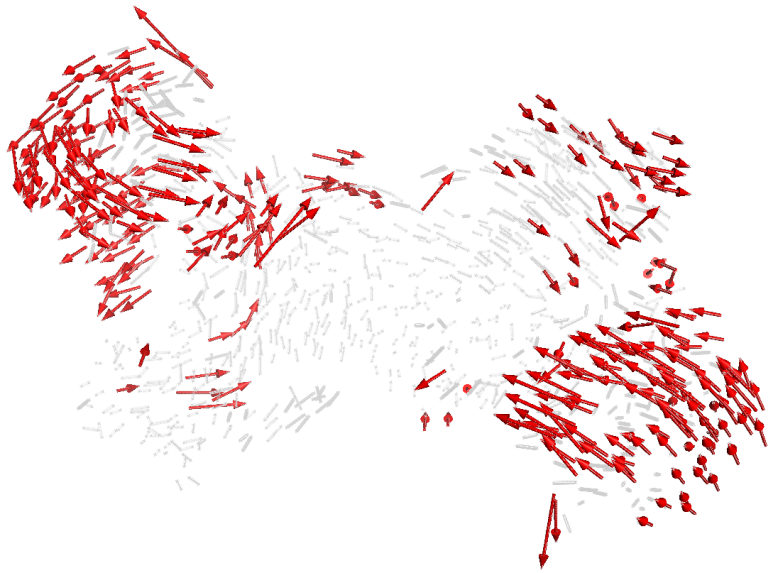


FIGURE A.5: Representation of the motions along one eigenvector visualised using PyMOL.

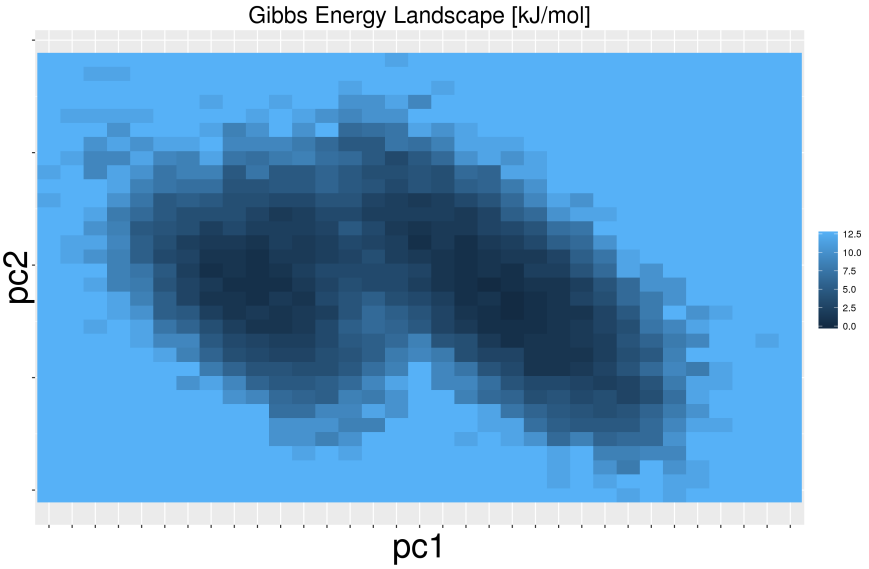


FIGURE A.6: Example of the PES profile obtained from the first two eigenvectors of a simulation.

A.5 Discussion

doitGROMACS was written to optimise the use of the different GROMACS tools in my workflow. Great efforts were made to rewrite the functions in a more general way, so that the use of doitGROMACS is possible for any protein system. Nevertheless there are some things that need to be improved. For example, the function that creates indexes of atoms (*indexCreator*()) is still very rudimentary. It requires the user to specify the exact syntax used by the GROMACS tools and it does not check for other user-defined indexes. It would also be ideal to reduce the user interactions by adding default values to some parameters, that can be overwritten, if needed, from the command line. However, the increase in functionalities and in complexity suggests a rethink of the role of doitGROMACS, a software written with simplicity in mind. One possible strategy would be the split into several scripts that do one single task each. The scripts will independently handle I/O files and error-control, and the new doitGROMACS would simply consist in a wrap that connects them together.

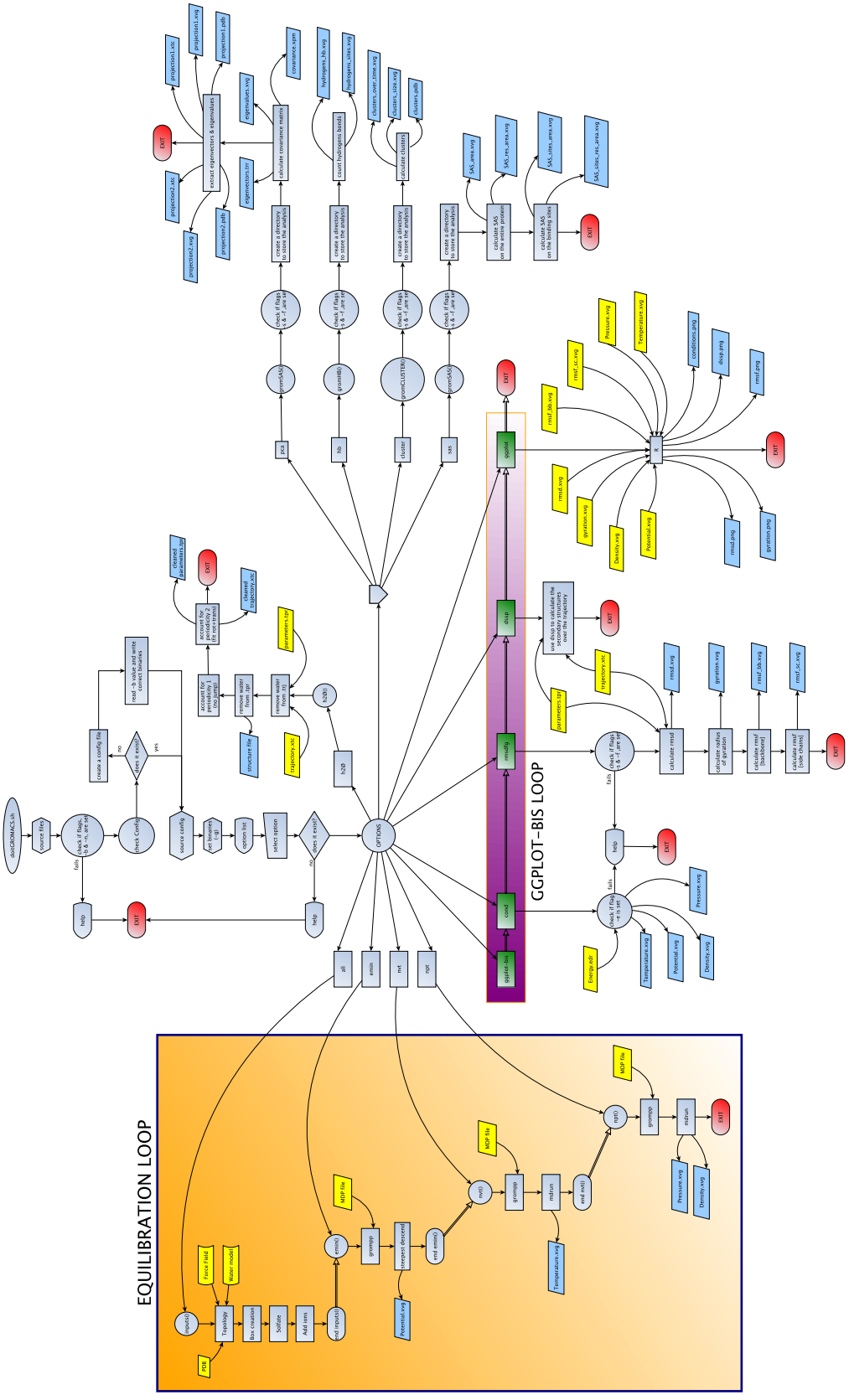


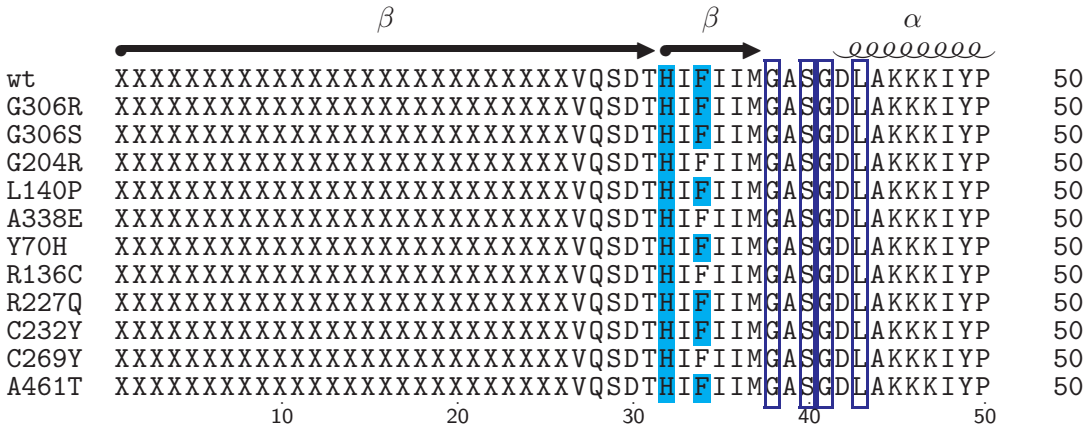
FIGURE A.7: Flowchart of doitgromacs.sh.

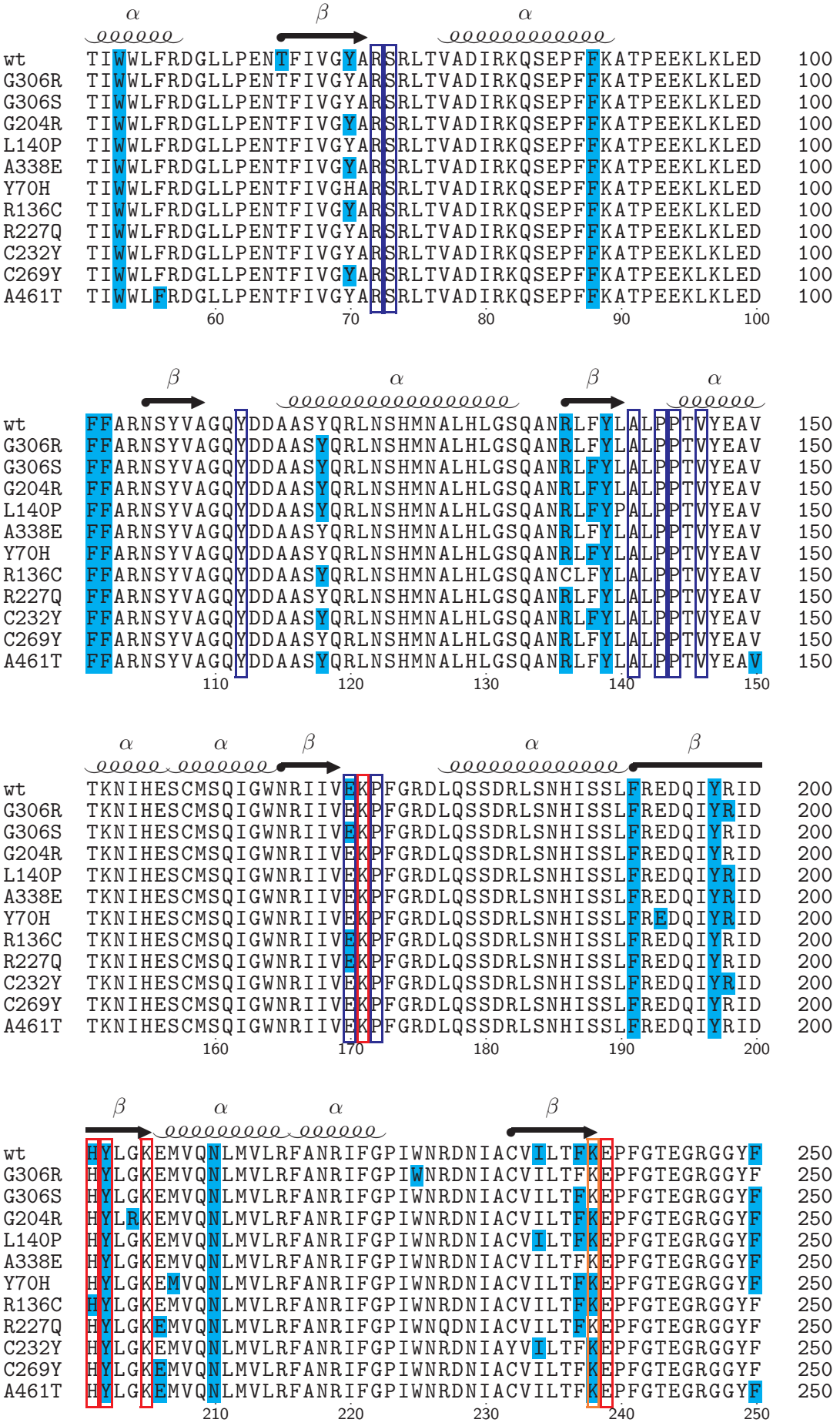
Appendix B

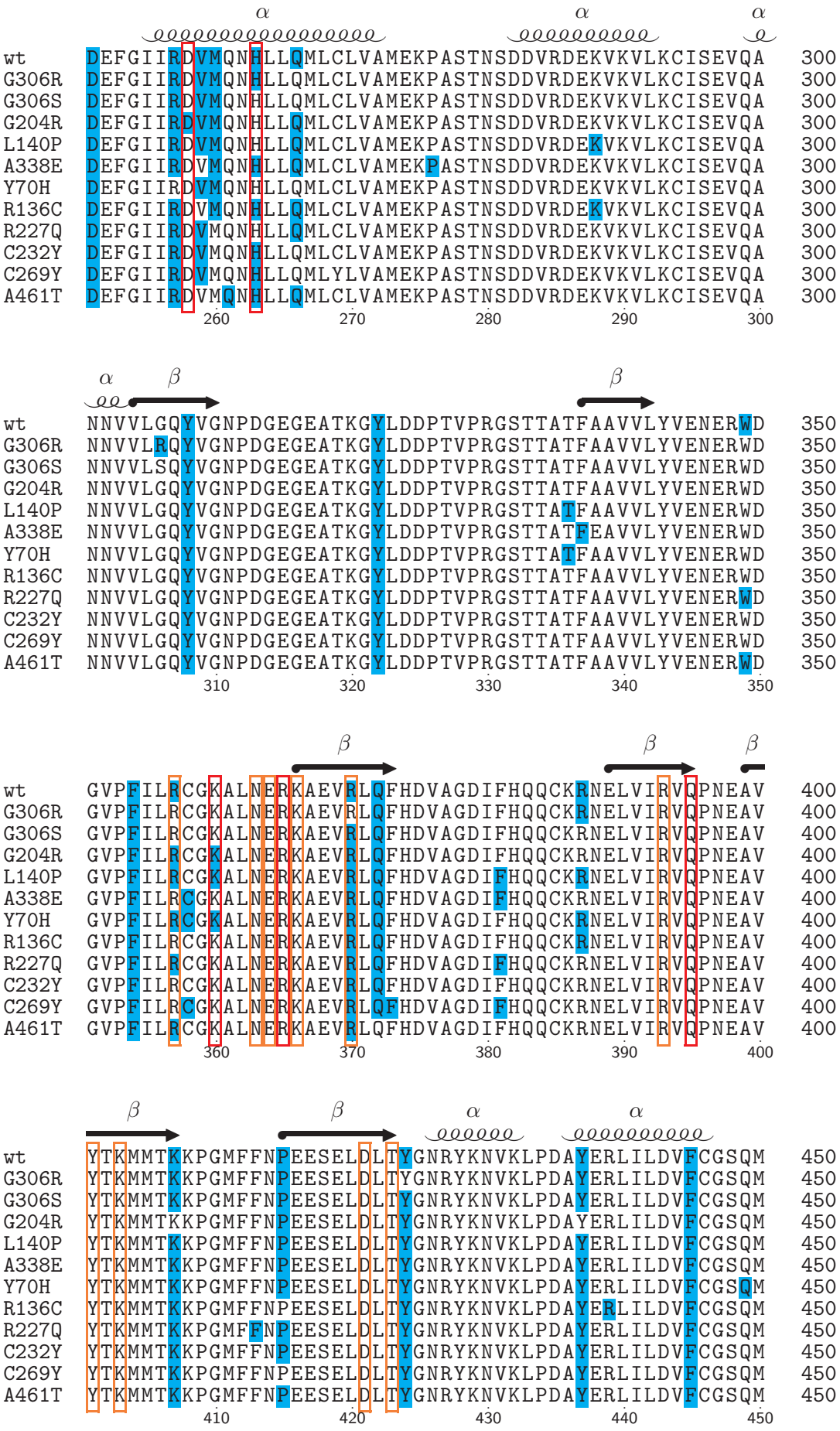
PSN alignments

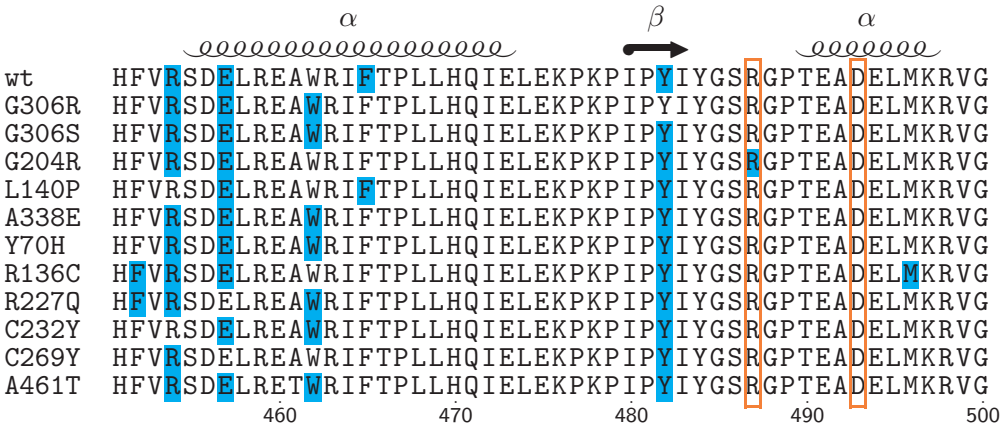
In the alignments, the hubs are marked in blue while the residues involved in binding are circled using different colour depending on the binding site: red for the G6P binding site, blue for the co-enzyme binding site and orange for the structural $NADP^+$ binding site.

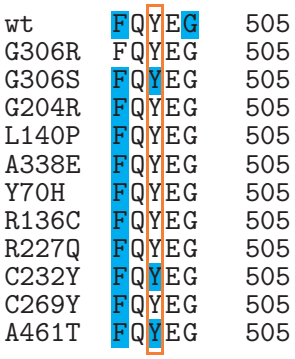
B.1 Hubs for the 310 K dynamics



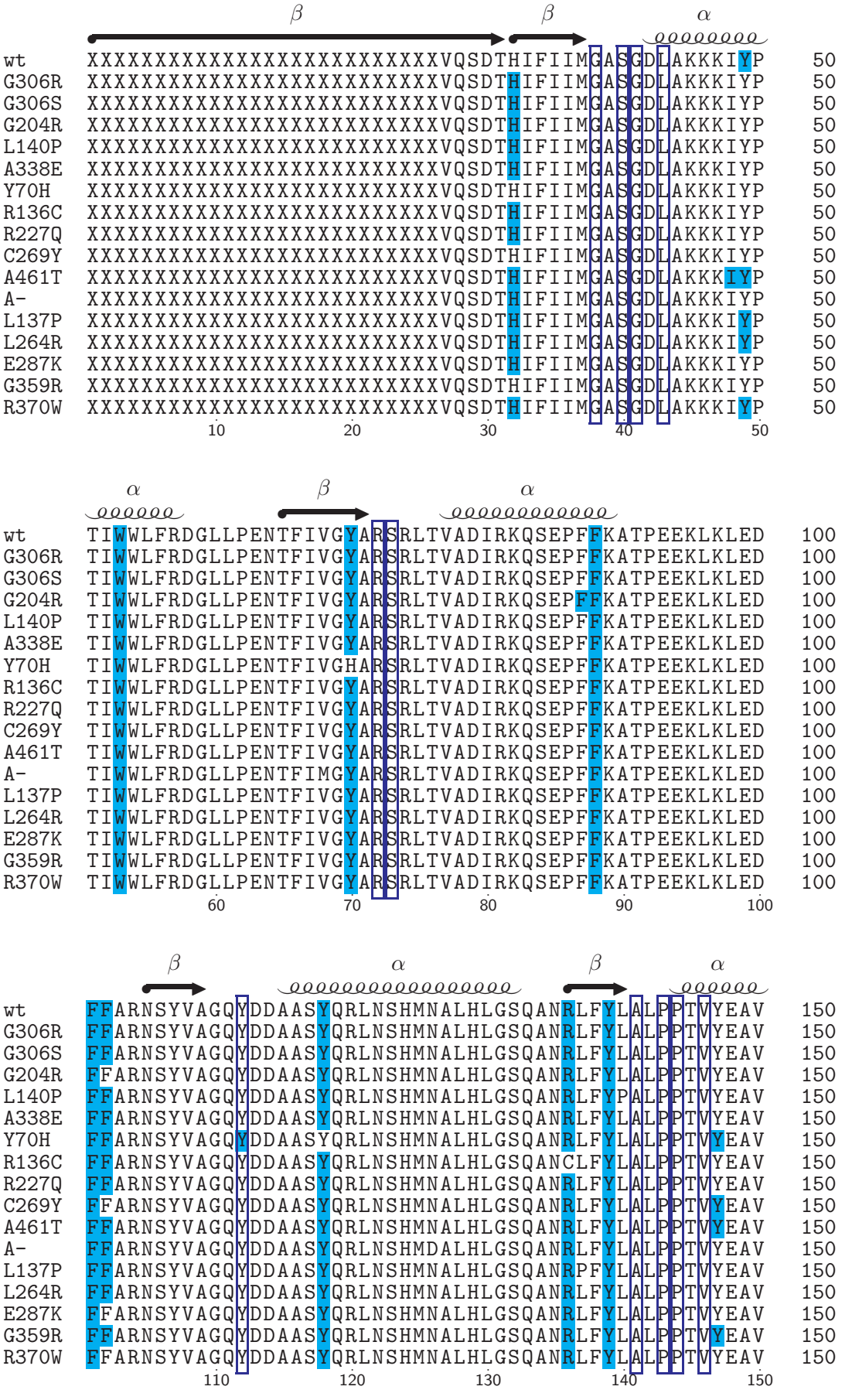

α β α
wt TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
G306R TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
G306S TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
G204R TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
L140P TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
A338E TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
Y70H TIWWLFRDGLLPENTFIVGHARSRLTVADIRKQSEPFFKATPEEKLKLED 100
R136C TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
R227Q TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
C232Y TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
C269Y TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
A461T TIWWLFRDGLLPENTFIVGYARSRLTVADIRKQSEPFFKATPEEKLKLED 100
60 70 80 90 100
β α β α
wt FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
G306R FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
G306S FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
G204R FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
L140P FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYPALPPTVYEAV 150
A338E FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
Y70H FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
R136C FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANCLFYLALPPTVYEAV 150
R227Q FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
C232Y FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
C269Y FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
A461T FFARNSYVAGQYDDAASYQRLNSHMNALHLGSQANRLFYLALPPTVYEAV 150
110 120 130 140 150
α α β α β
wt TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
G306R TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
G306S TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
G204R TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
L140P TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
A338E TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
Y70H TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
R136C TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
R227Q TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
C232Y TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
C269Y TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
A461T TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
160 170 180 190 200
β α α β
wt HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
G306R HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
G306S HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
G204R HYLRKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
L140P HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
A338E HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
Y70H HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
R136C HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
R227Q HYLGKEMVQNLMVLRFANRIFGPIWNQDNIACVILTFKEPFGTEGRGGYF 250
C232Y HYLGKEMVQNLMVLRFANRIFGPIWNRDNIAYVILTFKEPFGTEGRGGYF 250
C269Y HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
A461T HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
210 220 230 240 250

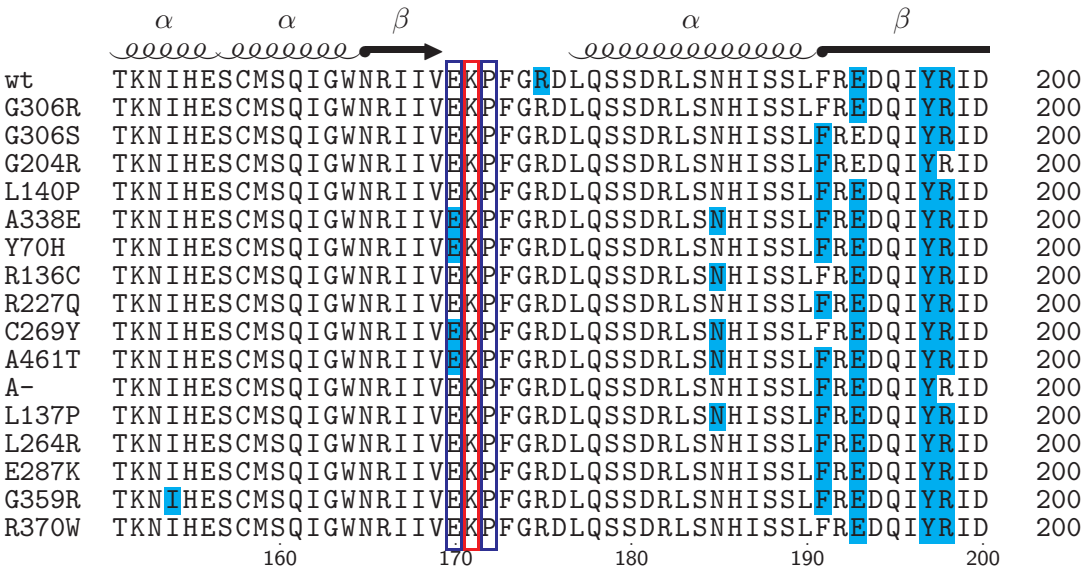

α α α
wt DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
G306R DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
G306S DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
G204R DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
L140P DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
A338E DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
Y70H DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
R136C DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
R227Q DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
C232Y DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
C269Y DEFGIIRDVMQNHLLQMLYLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
A461T DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
260 270 280 290 300
α β β
wt NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
G306R NNVVLRQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
G306S NNVVLSQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
G204R NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
L140P NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
A338E NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFEAVVLYVENERWD 350
Y70H NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
R136C NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
R227Q NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
C232Y NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
C269Y NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
A461T NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
310 320 330 340 350
β β β
wt GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
G306R GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
G306S GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
G204R GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
L140P GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
A338E GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
Y70H GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
R136C GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
R227Q GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
C232Y GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
C269Y GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
A461T GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
360 370 380 390 400
β β α α
wt YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
G306R YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
G306S YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
G204R YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
L140P YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
A338E YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
Y70H YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
R136C YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
R227Q YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
C232Y YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
C269Y YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
A461T YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
410 420 430 440 450

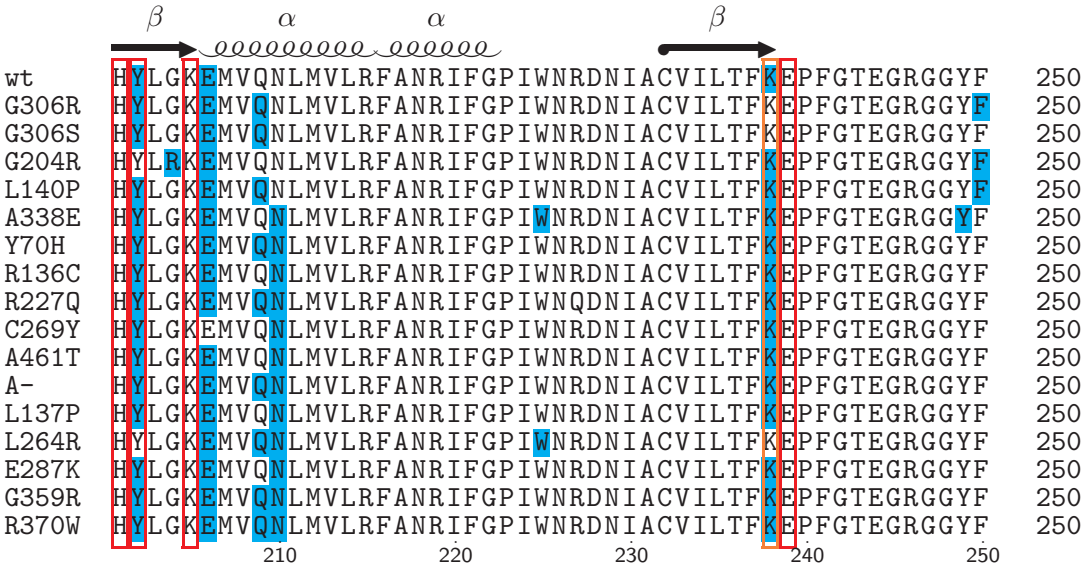

α
β
α
wt HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
G306R HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
G306S HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
G204R HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
L140P HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
A338E HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
Y70H HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
R136C HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
R227Q HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
C232Y HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
C269Y HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
A461T HFVRSDELRETWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
460
470
480
490
500


wt FQYEG 505
G306R FQYEG 505
G306S FQYEG 505
G204R FQYEG 505
L140P FQYEG 505
A338E FQYEG 505
Y70H FQYEG 505
R136C FQYEG 505
R227Q FQYEG 505
C232Y FQYEG 505
C269Y FQYEG 505
A461T FQYEG 505

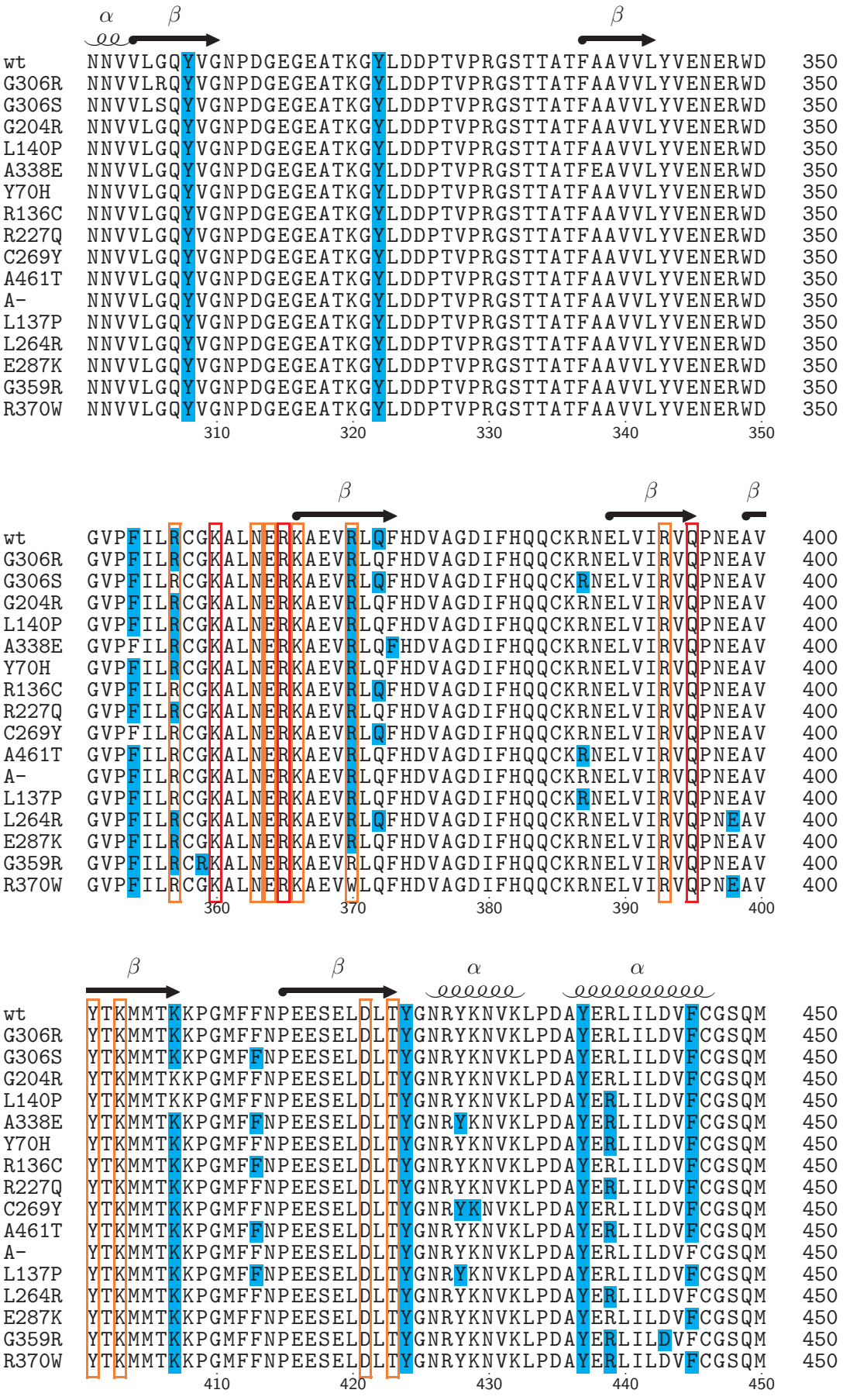
B.2 Hubs for the 400 K dynamics

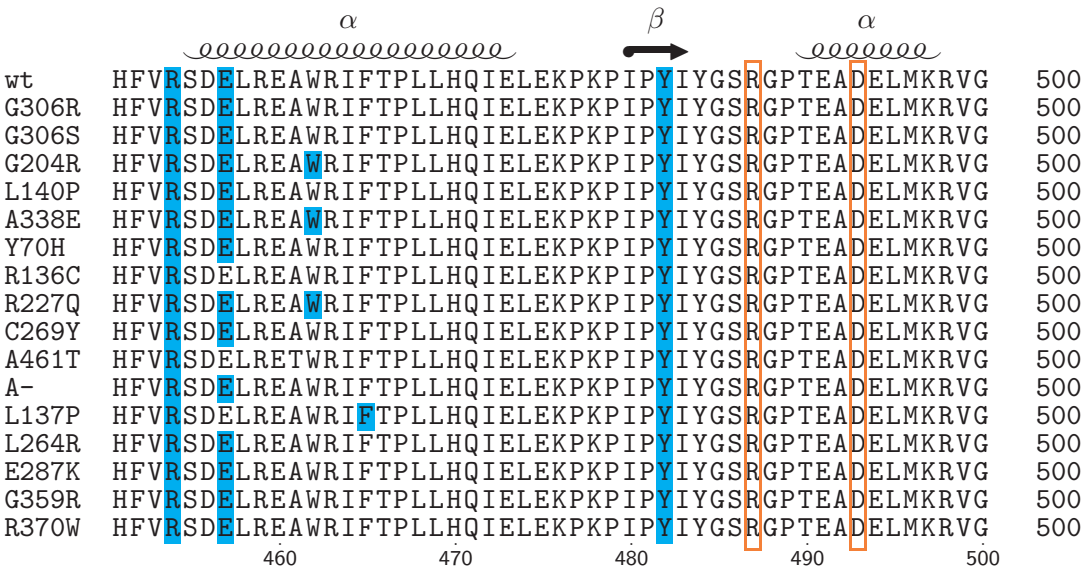


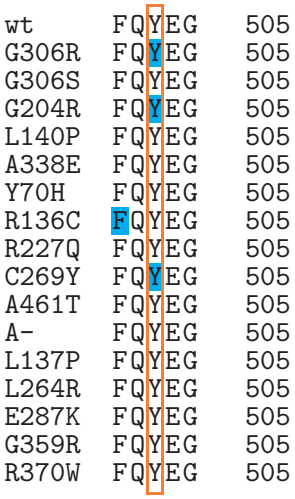

α α β α β
wt TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
G306R TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
G306S TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
G204R TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
L140P TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
A338E TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
Y70H TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
R136C TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
R227Q TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
C269Y TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
A461T TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
A- TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
L137P TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
L264R TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
E287K TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
G359R TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
R370W TKNIHESCMSQIGWNRIIVEKPFGRDLQSSDRLSNHISSLFREDQIYRID 200
160 170 180 190 200


β α α β
wt HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
G306R HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
G306S HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
G204R HYLRKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
L140P HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
A338E HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
Y70H HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
R136C HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
R227Q HYLGKEMVQNLMVLRFANRIFGPIWNQDNIACVILTFKEPFGTEGRGGYF 250
C269Y HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
A461T HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
A- HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
L137P HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
L264R HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
E287K HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
G359R HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
R370W HYLGKEMVQNLMVLRFANRIFGPIWNRDNIACVILTFKEPFGTEGRGGYF 250
210 220 230 240 250


α α α
wt DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
G306R DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
G306S DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
G204R DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
L140P DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
A338E DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
Y70H DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
R136C DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
R227Q DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
C269Y DEFGIIRDVMQNHLLQMLYLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
A461T DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
A- DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
L137P DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
L264R DEFGIIRDVMQNHRLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
E287K DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDKKVKVLKCISEVQA 300
G359R DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
R370W DEFGIIRDVMQNHLLQMLCLVAMEKPASTNSDDVRDEKVKVLKCISEVQA 300
260 270 280 290 300


α β β
wt NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
G306R NNVVLRQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
G306S NNVVLSQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
G204R NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
L140P NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
A338E NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFEAVVLYVENERWD 350
Y70H NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
R136C NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
R227Q NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
C269Y NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
A461T NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
A- NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
L137P NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
L264R NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
E287K NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
G359R NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
R370W NNVVLGQYVGNPDGEGEATKGYLDDPTVPRGSTTATFAAVVLYVENERWD 350
310 320 330 340 350
β β β
wt GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
G306R GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
G306S GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
G204R GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
L140P GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
A338E GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
Y70H GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
R136C GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
R227Q GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
C269Y GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
A461T GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
A- GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
L137P GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
L264R GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
E287K GVPFILRCGKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
G359R GVPFILRCRKALNERKAEVRLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
R370W GVPFILRCGKALNERKAEVWLQFHDVAGDIFHQQCKRNELVIRVQPNEAV 400
360 370 380 390 400
β β α α
wt YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
G306R YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
G306S YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
G204R YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
L140P YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
A338E YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
Y70H YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
R136C YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
R227Q YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
C269Y YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
A461T YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
A- YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
L137P YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
L264R YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
E287K YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
G359R YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
R370W YTKMMTKKPGMFFNPEESELDLTYGNRYKNVKLPDAYERLILDVFCGSQM 450
410 420 430 440 450


α
β
α
wt HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
G306R HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
G306S HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
G204R HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
L140P HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
A338E HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
Y70H HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
R136C HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
R227Q HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
C269Y HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
A461T HFVRSDELRETWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
A- HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
L137P HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
L264R HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
E287K HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
G359R HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
R370W HFVRSDELREAWRIFTPLLHQIELEKPKPIPYIYGSRGPTEADELMKRVG 500
460
470
480
490
500


wt FQYEG 505
G306R FQYEG 505
G306S FQYEG 505
G204R FQYEG 505
L140P FQYEG 505
A338E FQYEG 505
Y70H FQYEG 505
R136C FQYEG 505
R227Q FQYEG 505
C269Y FQYEG 505
A461T FQYEG 505
A- FQYEG 505
L137P FQYEG 505
L264R FQYEG 505
E287K FQYEG 505
G359R FQYEG 505
R370W FQYEG 505

Appendix C

All-atom tables

Tables containing the principal energies presented as averages over the trajectories.

C.1 Wild-type

TABLE C.1: The main measurements for the simulations performed on the wild-type as averages over the entire trajectories.

Mutant	T [K]	n°	time [ns]	E_{pot} [kJ/mol]	density [kg/m^3]	Rmsd [nm]	Gyration [nm]
Wild-type	310	1	250	-2.49×10^6	1002.9	0.22	3.65
		2	200	-2.49×10^6	1002.9	0.19	3.66
		3	250	-2.49×10^6	1002.9	0.27	3.69
			700	-2.49×10^6	1002.9	2.23	33.6
	400	1	1000	-1.99×10^6	906.3	0.38	3.72
		2	200	-1.99×10^6	906.5	0.43	3.77
			1200	-1.99×10^6	906.4	0.45	3.74
	450	1	500	-1.92×10^6	829.7	0.87	3.57
	500	1	200	-1.8×10^6	724.9	1.38	3.43
		2	200	-1.8×10^6	724.9	1.23	3.46
		3	200	-1.8×10^6	724.9	1.38	3.47
			600	-1.8×10^6	724.9	1.33	3.45
		9	3000				

TABLE C.2: The values (nm^2) of the area exposed to the solvent (SAS) as averages over the entire trajectories. All the values but SAS$_tot$-prob were calculated using a probe of standard size (0.14 nm), while for SAS*tot*-prob the probe was increased to 0.7 nm (see Section 3.3.5)

Mutant	T [K]	n°	SAS$_{tot}$-prob	SAS$_{tot}$	SAS$_{G6P}$	SAS$_{co-enzyme}$	SAS$_{struNADP^+}$
Wild-type	310	1	368.6	405.5	8.1	14.15	12.72
		2	370.3	405.3	7.87	13.9	13.43
		3	370.6	407.6	8.71	14.4	14.16
			369.8	406.1	8.22	14.1	13.4
	400	1	375.1	407.3	8.48	14.47	12.82
		2	375.9	408.8	9.25	14.84	12.09
			375.5	408	8.86	14.6	12.4
	450	1	371.8	428	5.94	10.66	13.54
	500	1	3723	458.4	5.70	11.73	11.6
		2	375.1	451.5	5.22	11.5	12.6
		3	377.4	459.6	6.63	15.15	12.06
			374.9	456.5	5.85	12.8	12.1

C.2 Mutants

TABLE C.3: The main measurements for the simulations performed on the mutants as averages over the entire trajectories.

Mutant	T [K]	n°	time [ns]	E_{pot} [kJ/mol]	density [kg/m^3]	Rmsd [nm]	Gyration [nm]
G306R	310	1	200	-2.39×10^6	1004	0.24	3.69
	400	1	500	-2.1×10^6	904	0.37	3.7
	450	1	500	-1.92×10^6	829.7	0.8	3.75
		3	1200				
G306S	310	1	200	-2.49×10^6	1003	0.2	3.66
		2	100	-2.49×10^6	1003	0.2	3.67
			300	-2.49×10^6	1003	0.2	3.66
	400	1	200	-2.1×10^6	904.8	0.3	3.71
		2	200	-2.1×10^6	904.9	0.3	3.68
			400	-2.1×10^6	904.9	0.3	3.69
		4	700				
A$^-$	400	1	500	-2.18×10^6	903.7	0.32	3.67
	450	1	500	-2.0×10^6	828.3	0.68	3.72
	470	1	500	-1.76×10^6	795	1.04	3.69
		3	1500				
Y70H	310	1	150	-2.49×10^6	1002	0.3	3.67
	400	1	200	-1.99×10^6	906.4	0.42	3.69
		2	350				
R136C	310	1	150	-2.48×10^6	1002.9	0.23	3.68
	400	1	200	-2.1×10^6	904.7	0.32	3.7
		2	200	-2.1×10^6	904.8	0.36	3.7
		3	550				
G204R	310	1	200	-2.39×10^6	1004	0.3	3.73
	400	1	200	-1.99×10^6	906.4	0.4	3.74
		2	400				

Mutant	T [K]	n°	time [ns]	E_{pot} [kJ/mol]	density [kg/m^3]	Rmsd [nm]	Gyration [nm]
A461T	310	1	150	-2.48×10^6	1002.9	0.21	3.67
	400	1	200	-2.1×10^6	904.8	0.45	3.74
		2	350				
R227Q	310	1	150	-2.49×10^6	1002.9	0.24	3.69
	400	1	200	-1.99×10^6	906.5	0.33	3.67
		2	350				
A338E	310	1	200	-2.39×10^6	1004	0.26	3.68
	400	1	500	-2.1×10^6	904.8	0.27	3.68
	450	1	500	-1.84×10^6	831.4	0.68	3.69
		3	1200				
L140P	310	1	200	-2.39×10^6	1004	0.24	3.69
	400	1	500	-2.1×10^6	904.8	0.32	3.66
		2	700				
C269Y	310	1	150	-2.48×10^6	1002.9	0.26	3.68
	400	1	200	-1.99×10^6	906.5	0.37	3.72
		2	350				

TABLE C.4: The values (nm^2) of the area exposed to the solvent (SAS) as averages over the entire trajectories. All the values but SAS$_tot$-prob were calculated using a probe of standard size (0.14 nm), while for SAS*tot*-prob the probe was increased to 0.7 nm (see Section 3.3.5)

Mutant	T [K]	n°	SAS$_{tot}$-prob	SAS$_{tot}$	SAS$_{G6P}$	SAS$_{co-enzyme}$	SAS$_{struNADP^+}$
G306R	310	1	371.2	409.9	7.6	13.8	14.8
	400	1	367.9	404.5	7.2	13.6	13.3
	450	1	384.1	436.3	5.2	10.83	12.3
G306S	310	1	373.2	409.6	8.8	15.2	12.4
		2	369.6	407.4	8.4	14.0	13.6
			471.4	408.5	8.6	14.6	13
	400	1	373.7	405.6	8.2	14.7	12.8
		2	371.6	404.3	7.5	14.3	12.8
			372.4	404.9	7.8	14.5	12.8
G204R	310	1	374.1	409.5	8.3	14.4	12.7
	400	1	377.5	407.3	7.7	13.9	11.8
C269Y	310	1	372.4	409.9	8.0	13.3	13.4
	400	1	373.2	405.7	8.1	14.9	12.7
R227Q	310	1	372.2	405.3	7.9	14.8	13.0
	400	1	367.7	400.9	7.1	13.9	11.5
A461T	310	1	369.9	403.3	8.1	14.2	12.6
	400	1	377.8	407.9	9.2	14.7	13.0
R136C	310	1	369.7	410.6	8.4	13.9	12.7
	400	1	372.3	406.7	7.4	13.4	12.6
		2	372.3	401.9	8.6	15.3	12
A^-	400	1	373.5	404.5	6.8	14.1	11.2
	450	1	373	419.8	4.8	9	12.9
	470	1	371.2	431.7	4.1	9.4	10.9
A338E	310	1	375.7	408.9	7.5	14.6	12.9
	400	1	376.7	408.9	8.7	14.7	13.1
	450	1	373.6	420.8	4.56	10.9	12.9
L140P	310	1	372.5	406.8	8.6	15.2	12.0
	400	1	367.4	395.8	6.1	13.8	12.1
3Y70H	310	1	376.9	410.2	8.9	15.8	12.5
	400	1		405.5	8.2	14.9	12.5

Bibliography

[1] J. K. Baird, "Neglect of Plasmodium vivax malaria," *Trends Parasitol.*, vol. 23, pp. 533–539, Nov 2007.

[2] R. N. Price, E. Tjitra, C. A. Guerra, S. Yeung, N. J. White, and N. M. Anstey, "Vivax malaria: neglected and not benign," *Am. J. Trop. Med. Hyg.*, vol. 77, pp. 79–87, Dec 2007.

[3] J. Field and P. Shute, "Plasmodium vivax. in the microscopuic diagnosis of human malaria. a morphological study of the erythrocytic parasites.," *Institute for Medical Research*, vol. 2, 1956.

[4] M. Karyana, L. Burdarm, S. Yeung, E. Kenangalem, N. Wariker, R. Maristela, K. G. Umana, R. Vemuri, M. J. Okoseray, P. M. Penttinen, P. Ebsworth, P. Sugiarto, N. M. Anstey, E. Tjitra, and R. N. Price, "Malaria morbidity in Papua Indonesia, an area with multidrug resistant Plasmodium vivax and Plasmodium falciparum," *Malar. J.*, vol. 7, p. 148, 2008.

[5] C. J. Hemmer, F. G. Holst, P. Kern, C. B. Chiwakata, M. Dietrich, and E. C. Reisinger, "Stronger host response per parasitized erythrocyte in Plasmodium vivax or ovale than in Plasmodium falciparum malaria," *Trop. Med. Int. Health*, vol. 11, pp. 817–823, Jun 2006.

[6] N. D. Karunaweera, G. E. Grau, P. Gamage, R. Carter, and K. N. Mendis, "Dynamics of fever and serum levels of tumor necrosis factor are closely associated during clinical paroxysms in Plasmodium vivax malaria," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 89, pp. 3200–3203, Apr 1992.

[7] R. Suwanarusk, B. M. Cooke, A. M. Dondorp, K. Silamut, J. Sattabongkot, N. J. White, and R. Udomsangpetch, "The deformability of red blood cells parasitized by Plasmodium falciparum and P. vivax," *J. Infect. Dis.*, vol. 189, pp. 190–194, Jan 2004.

[8] S. Handayani, D. T. Chiu, E. Tjitra, J. S. Kuo, D. Lampah, E. Kenangalem, L. Renia, G. Snounou, R. N. Price, N. M. Anstey, and B. Russell, "High deformability

of Plasmodium vivax-infected red blood cells under microfluidic conditions," *J. Infect. Dis.*, vol. 199, pp. 445–450, Feb 2009.

[9] J. K. Baird, E. Schwartz, and S. L. Hoffman, "Prevention and treatment of vivax malaria," *Curr Infect Dis Rep*, vol. 9, pp. 39–46, Jan 2007.

[10] C. S. Boutlis, T. W. Yeo, and N. M. Anstey, "Malaria tolerance–for whom the cell tolls?," *Trends Parasitol.*, vol. 22, pp. 371–377, Aug 2006.

[11] E. Tjitra, N. M. Anstey, P. Sugiarto, N. Warikar, E. Kenangalem, M. Karyana, D. A. Lampah, and R. N. Price, "Multidrug-resistant Plasmodium vivax associated with severe and fatal malaria: a prospective study in Papua, Indonesia," *PLoS Med.*, vol. 5, p. e128, Jun 2008.

[12] J. J. Lipa, "Twentieth report from the WHO Expert Commitee on Malaria. WHO technical report No. 892, Geneva," *Wiad Parazytol*, vol. 46, no. 4, pp. 523–524, 2000.

[13] R. Feachem and O. Sabot, "A new global malaria eradication strategy," *Lancet*, vol. 371, pp. 1633–1635, May 2008.

[14] B. Sina, "Focus on Plasmodium vivax," *Trends Parasitol.*, vol. 18, pp. 287–289, Jul 2002.

[15] J. K. Baird, "Chloroquine resistance in Plasmodium vivax," *Antimicrob. Agents Chemother.*, vol. 48, pp. 4075–4083, Nov 2004.

[16] M. Arevalo-Herrera, C. Chitnis, and S. Herrera, "Current status of Plasmodium vivax vaccine," *Hum Vaccin*, vol. 6, pp. 124–132, Jan 2010.

[17] J. M. Carlton, A. A. Escalante, D. Neafsey, and S. K. Volkman, "Comparative evolutionary genomics of human malaria parasites," *Trends Parasitol.*, vol. 24, pp. 545–550, Dec 2008.

[18] K. C. Kain, A. E. Brown, H. K. Webster, R. A. Wirtz, J. S. Keystone, M. H. Rodriguez, J. Kinahan, M. Rowland, and D. E. Lanar, "Circumsporozoite genotyping of global isolates of Plasmodium vivax from dried blood specimens," *J. Clin. Microbiol.*, vol. 30, pp. 1863–1866, Jul 1992.

[19] R. Tewari, R. Spaccapelo, F. Bistoni, A. A. Holder, and A. Crisanti, "Function of region I and II adhesive motifs of Plasmodium falciparum circumsporozoite protein in sporozoite motility and infectivity," *J. Biol. Chem.*, vol. 277, pp. 47613–47618, Dec 2002.

[20] A. Coppi, C. Pinzon-Ortiz, C. Hutter, and P. Sinnis, "The Plasmodium circumsporozoite protein is proteolytically processed during cell invasion," *J. Exp. Med.*, vol. 201, pp. 27–33, Jan 2005.

[21] M. Yuda and T. Ishino, "Liver invasion by malarial parasites–how do malarial parasites break through the host barrier?," *Cell. Microbiol.*, vol. 6, pp. 1119–1125, Dec 2004.

[22] W. O. Rogers, K. Gowda, and S. L. Hoffman, "Construction and immunogenicity of DNA vaccine plasmids encoding four Plasmodium vivax candidate vaccine antigens," *Vaccine*, vol. 17, pp. 3136–3144, Aug 1999.

[23] A. Castellanos, M. Arevalo-Herrera, N. Restrepo, L. Gulloso, G. Corradin, and S. Herrera, "Plasmodium vivax thrombospondin related adhesion protein: immunogenicity and protective efficacy in rodents and Aotus monkeys," *Mem. Inst. Oswaldo Cruz*, vol. 102, pp. 411–416, Jun 2007.

[24] J. H. Adams, D. E. Hudson, M. Torii, G. E. Ward, T. E. Wellems, M. Aikawa, and L. H. Miller, "The Duffy receptor family of Plasmodium knowlesi is located within the micronemes of invasive malaria merozoites," *Cell*, vol. 63, pp. 141–153, Oct 1990.

[25] J. Cole-Tobian and C. L. King, "Diversity and natural selection in Plasmodium vivax Duffy binding protein gene," *Mol. Biochem. Parasitol.*, vol. 127, pp. 121–132, Apr 2003.

[26] E. M. Riley, S. J. Allen, J. G. Wheeler, M. J. Blackman, S. Bennett, B. Takacs, H. J. Schonfeld, A. A. Holder, and B. M. Greenwood, "Naturally acquired cellular and humoral immune responses to the major merozoite surface antigen (PfMSP1) of Plasmodium falciparum are associated with reduced malaria morbidity," *Parasite Immunol.*, vol. 14, pp. 321–337, May 1992.

[27] M. A. Herrera, F. Rosero, S. Herrera, P. Caspers, D. Rotmann, F. Sinigaglia, and U. Certa, "Protection against malaria in Aotus monkeys immunized with a recombinant blood-stage antigen fused to a universal T-cell epitope: correlation of serum gamma interferon levels with protection," *Infect. Immun.*, vol. 60, pp. 154–158, Jan 1992.

[28] P. Graves and H. Gelband, "Vaccines for preventing malaria (blood-stage)," *Cochrane Database Syst Rev*, no. 4, p. CD006199, 2006.

[29] D. L. Doolan and S. L. Hoffman, "Multi-gene vaccination against malaria: A multistage, multi-immune response approach," *Parasitol. Today (Regul. Ed.)*, vol. 13, pp. 171–178, May 1997.

[30] G. H. Mitchell, A. W. Thomas, G. Margos, A. R. Dluzewski, and L. H. Bannister, "Apical membrane antigen 1, a major malaria vaccine candidate, mediates the close attachment of invasive merozoites to host red blood cells," *Infect. Immun.*, vol. 72, pp. 154–158, Jan 2004.

[31] M. H. Rodrigues, K. M. Rodrigues, T. R. Oliveira, A. N. Comodo, M. M. Rodrigues, C. H. Kocken, A. W. Thomas, and I. S. Soares, "Antibody response of naturally infected individuals to recombinant Plasmodium vivax apical membrane antigen-1," *Int. J. Parasitol.*, vol. 35, pp. 185–192, Feb 2005.

[32] H. Hisaeda, A. W. Stowers, T. Tsuboi, W. E. Collins, J. S. Sattabongkot, N. Suwanabun, M. Torii, and D. C. Kaslow, "Antibodies to malaria vaccine candidates Pvs25 and Pvs28 completely block the ability of Plasmodium vivax to infect mosquitoes," *Infect. Immun.*, vol. 68, pp. 6618–6623, Dec 2000.

[33] T. Tsuboi, D. C. Kaslow, M. M. Gozar, M. Tachibana, Y. M. Cao, and M. Torii, "Sequence polymorphism in two novel Plasmodium vivax ookinete surface proteins, Pvs25 and Pvs28, that are malaria transmission-blocking vaccine candidates," *Mol. Med.*, vol. 4, pp. 772–782, Dec 1998.

[34] T. N. Wells, J. N. Burrows, and J. K. Baird, "Targeting the hypnozoite reservoir of Plasmodium vivax: the hidden obstacle to malaria elimination," *Trends Parasitol.*, vol. 26, pp. 145–151, Mar 2010.

[35] D. R. Hill, J. K. Baird, M. E. Parise, L. S. Lewis, E. T. Ryan, and A. J. Magill, "Primaquine: report from CDC expert meeting on malaria chemoprophylaxis I," *Am. J. Trop. Med. Hyg.*, vol. 75, pp. 402–415, Sep 2006.

[36] J. Soto, J. Toledo, M. Rodriquez, J. Sanchez, R. Herrera, J. Padilla, and J. Berman, "Double-blind, randomized, placebo-controlled assessment of chloroquine/primaquine prophylaxis for malaria in nonimmune Colombian soldiers," *Clin. Infect. Dis.*, vol. 29, pp. 199–201, Jul 1999.

[37] E. Schwartz and G. Regev-Yochay, "Primaquine as prophylaxis for malaria for nonimmune travelers: A comparison with mefloquine and doxycycline," *Clin. Infect. Dis.*, vol. 29, pp. 1502–1506, Dec 1999.

[38] L. Luzzatto and G. Battistuzzi, "Glucose-6-phosphate dehydrogenase," *Adv. Hum. Genet.*, vol. 14, pp. 217–329, 1985.

[39] L. Luzzatto and U. Testa, "Human erythrocyte glucose 6-phosphate dehydrogenase: structure and function in normal and mutant subjects," *Curr Top Hematol*, vol. 1, pp. 1–70, 1978.

[40] J. K. Baird and K. H. Rieckmann, "Can primaquine therapy for vivax malaria be improved?," *Trends Parasitol.*, vol. 19, pp. 115–120, Mar 2003.

[41] J. L. Goller, D. Jolley, P. Ringwald, and B. A. Biggs, "Regional differences in the response of Plasmodium vivax malaria to primaquine as anti-relapse therapy," *Am. J. Trop. Med. Hyg.*, vol. 76, pp. 203–207, Feb 2007.

[42] N. J. Elmes, P. E. Nasveld, S. J. Kitchener, D. A. Kocisko, and M. D. Edstein, "The efficacy and tolerability of three different regimens of tafenoquine versus primaquine for post-exposure prophylaxis of Plasmodium vivax malaria in the Southwest Pacific," *Trans. R. Soc. Trop. Med. Hyg.*, vol. 102, pp. 1095–1101, Nov 2008.

[43] P. Cohen and M. A. Rosemeyer, "Subunit interactions of glucose-6-phosphate dehydrogenase from human erythrocytes," *Eur. J. Biochem.*, vol. 8, pp. 8–15, Mar 1969.

[44] P. Cohen and M. A. Rosemeyer, "Subunit interactions of glucose-6-phosphate dehydrogenase from human erythrocytes," *European Journal of Biochemistry*, vol. 8, no. 1, pp. 8–15, 1969.

[45] H.-G. Zimmer, *Pentose Phosphate Pathway.* John Wiley & Sons, Ltd, 2001.

[46] J. E. Smith and E. Beutler, "Anomeric specificity of human erythrocyte glucose-6-phosphate dehydrogenase," *Proc. Soc. Exp. Biol. Med.*, vol. 122, pp. 671–673, Jul 1966.

[47] W. Scientific Group, "Standardization of procedures for the study of glucose-6-phosphate dehydrogenase. Report of a WHO Scientific Group," *World Health Organ Tech Rep Ser*, vol. 366, pp. 1–53, 1967.

[48] P. J. Mason, J. M. Bautista, and F. Gilsanz, "G6PD deficiency: the genotype-phenotype association," *Blood Rev*, vol. 21, pp. 267–283, 2007.

[49] M. J. Maisels, "Neonatal jaundice," *Pediatr Rev*, vol. 27, pp. 443–454, Dec 2006.

[50] Y. H. Weng, Y. H. Chou, and R. I. Lien, "Hyperbilirubinemia in healthy neonates with glucose-6-phosphate dehydrogenase deficiency," *Early Hum. Dev.*, vol. 71, pp. 129–136, Apr 2003.

[51] U. Bienzle, O. Sodeinde, C. E. Effiong, and L. Luzzatto, "Glucose 6-phosphate dehydrogenase deficiency and sickle cell anemia: frequency and features of the association in an African community," *Blood*, vol. 46, pp. 591–597, Oct 1975.

[52] S. C. Bernstein, J. E. Bowman, and L. K. Noche, "Interaction of sickle cell trait and glucose-6-phosphate dehydrogenase deficiency in Cameroon," *Hum. Hered.*, vol. 30, no. 1, pp. 7–11, 1980.

[53] A. Tagarelli, A. Piro, L. Bastone, and G. Tagarelli, "Identification of glucose 6-phosphate dehydrogenase deficiency in a population with a high frequency of thalassemia," *FEBS Lett.*, vol. 466, pp. 139–142, Jan 2000.

[54] M. H. Steinberg, M. S. West, D. Gallagher, and W. Mentzer, "Effects of glucose-6-phosphate dehydrogenase deficiency upon sickle cell anemia," *Blood*, vol. 71, pp. 748–752, Mar 1988.

[55] A. M. Than, T. Harano, K. Harano, A. A. Myint, T. Ogino, and S. Okadaa, "High incidence of 3-thalassemia, hemoglobin E, and glucose-6-phosphate dehydrogenase deficiency in populations of malaria-endemic southern Shan State, Myanmar," *Int. J. Hematol.*, vol. 82, pp. 119–123, Aug 2005.

[56] J. L. Vives-Corrons, W. Kuhl, M. A. Pujades, and E. Beutler, "Molecular genetics of the glucose-6-phosphate dehydrogenase (G6PD) Mediterranean variant and description of a new G6PD mutant, G6PD Andalus1361A," *Am. J. Hum. Genet.*, vol. 47, pp. 575–579, Sep 1990.

[57] T. J. Vulliamy, A. Othman, M. Town, A. Nathwani, A. G. Falusi, P. J. Mason, and L. Luzzatto, "Polymorphic sites in the African population detected by sequence analysis of the glucose-6-phosphate dehydrogenase gene outline the evolution of the variants A and A-," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 88, pp. 8568–8571, Oct 1991.

[58] D. T. Chiu, L. Zuo, E. Chen, L. Chao, E. Louie, B. Lubin, T. Z. Liu, and C. S. Du, "Two commonly occurring nucleotide base substitutions in Chinese G6PD variants," *Biochem. Biophys. Res. Commun.*, vol. 180, pp. 988–993, Oct 1991.

[59] A. C. Allison, "Glucose-6-phosphate dehydrogenase deficiency in red blood cells of East Africans," *Nature*, vol. 186, pp. 531–532, May 1960.

[60] A. G. Motulsky, "Metabolic polymorphisms and the role of infectious diseases in human evolution," *Hum. Biol.*, vol. 32, pp. 28–62, Feb 1960.

[61] M. Cappadoro, G. Giribaldi, E. O'Brien, F. Turrini, F. Mannu, D. Ulliers, G. Simula, L. Luzzatto, and P. Arese, "Early phagocytosis of glucose-6-phosphate dehydrogenase (G6PD)-deficient erythrocytes parasitized by Plasmodium falciparum may explain malaria protection in G6PD deficiency," *Blood*, vol. 92, pp. 2527–2534, Oct 1998.

[62] P. A. Marks and R. T. Gross, "Erythrocyte glucose-6-phosphate dehydrogenase deficiency: evidence of differences between Negroes and Caucasians with respect to this genetically determined trait," *J. Clin. Invest.*, vol. 38, pp. 2253–2262, Dec 1959.

[63] P. J. Mason, M. F. Sonati, D. MacDonald, C. Lanza, D. Busutil, M. Town, C. M. Corcoran, J. S. Kaeda, D. J. Stevens, and S. al Ismail, "New glucose-6-phosphate dehydrogenase mutations associated with chronic anemia," *Blood*, vol. 85, pp. 1377–1380, Mar 1995.

[64] L. Luzzatto and N. C. Allan, "Different properties of glucose 6-phosphate dehydrogenase from human erythrocytes with normal and abnormal enzyme levels," *Biochem. Biophys. Res. Commun.*, vol. 21, pp. 547–554, Dec 1965.

[65] H. N. Kirkman and E. M. Hendrickson, "Glucose 6-phosphate dehydrogenase from human erythrocytes. II. Subactive states of the enzyme from normal persons," *J. Biol. Chem.*, vol. 237, pp. 2371–2376, Jul 1962.

[66] L. Longo, O. C. Vanegas, M. Patel, V. Rosti, H. Li, J. Waka, T. Merghoub, P. P. Pandolfi, R. Notaro, K. Manova, and L. Luzzatto, "Maternally transmitted severe glucose 6-phosphate dehydrogenase deficiency is an embryonic lethal," *EMBO J.*, vol. 21, pp. 4229–4239, Aug 2002.

[67] P. E. Meissner, B. Coulibaly, G. Mandi, U. Mansmann, S. Witte, W. Schiek, O. Muller, R. H. Schirmer, F. P. Mockenhaupt, and U. Bienzle, "Diagnosis of red cell G6PD deficiency in rural Burkina Faso: comparison of a rapid fluorescent enzyme test on filter paper with polymerase chain reaction based genotyping," *Br. J. Haematol.*, vol. 131, pp. 395–399, Nov 2005.

[68] N. Carter, A. Pamba, S. Duparc, and J. N. Waitumbi, "Frequency of glucose-6-phosphate dehydrogenase deficiency in malaria patients from six African countries enrolled in two randomized anti-malarial clinical trials," *Malar. J.*, vol. 10, p. 241, 2011.

[69] G. J. Brewer, A. R. Tarlov, and A. S. Alving, "Methaemoglobin reduction test: a new, simple, in vitro test for identifying primaquine-sensitivity," *Bull. World Health Organ.*, vol. 22, pp. 633–640, 1960.

[70] B. Keats, "Genetic mapping: X chromosome," *Hum. Genet.*, vol. 64, no. 1, pp. 28–32, 1983.

[71] F. Kiani, S. Schwarzl, S. Fischer, and T. Efferth, "Three-dimensional modeling of glucose-6-phosphate dehydrogenase-deficient variants from German ancestry," *PLoS One*, vol. 2, pp. e625–e625, 2007.

[72] N. G. Wrigley, J. V. Heather, A. Bonsignore, and A. De Flora, "Human erythrocyte glucose 6-phosphate dehydrogenase: electron microscope studies on structure and interconversion of tetramers, dimers and monomers," *J. Mol. Biol.*, vol. 68, pp. 483–499, Jul 1972.

[73] S. W. Au, S. Gover, V. M. Lam, and M. J. Adams, "Human glucose-6-phosphate dehydrogenase: the crystal structure reveals a structural NADP(+) molecule and provides insights into enzyme deficiency," *Structure*, vol. 8, pp. 293–303, Mar 2000.

[74] D. J. Stevens, W. Wanachiwanawin, P. J. Mason, T. J. Vulliamy, and L. Luzzatto, "G6PD Canton a common deficient variant in South East Asia caused by a 459 Arg—-Leu mutation," *Nucleic Acids Res.*, vol. 18, p. 7190, Dec 1990.

[75] R. Notaro, A. Afolayan, and L. Luzzatto, "Human mutations in glucose 6-phosphate dehydrogenase reflect evolutionary history," *FASEB J.*, vol. 14, pp. 485–494, Mar 2000.

[76] Y. S. Cheng, T. K. Tang, and M. Hwang, "Amino acid conservation and clinical severity of human glucose-6-phosphate dehydrogenase mutations," *J. Biomed. Sci.*, vol. 6, no. 2, pp. 106–114, 1999.

[77] S. W. N. Au, C. E. Naylor, S. Gover, L. Vandeputte-Rutten, D. A. Scopes, P. J. Mason, L. Luzzatto, V. M. S. Lam, and M. J. Adams, "Solution of the structure of tetrameric human glucose 6-phosphate dehydrogenase by molecular replacement," *Acta Crystallographica Section D*, vol. 55, pp. 826–834, Apr 1999.

[78] L. Camardella, C. Caruso, B. Rutigliano, M. Romano, G. Di Prisco, and F. Descalzi-Cancedda, "Human erythrocyte glucose-6-phosphate dehydrogenase. Identification of a reactive lysyl residue labelled with pyridoxal 5'-phosphate," *Eur. J. Biochem.*, vol. 171, pp. 485–489, Feb 1988.

[79] W. T. Lee and H. R. Levy, "Lysine-21 of Leuconostoc mesenteroides glucose 6-phosphate dehydrogenase participates in substrate binding through charge-charge interaction," *Protein Sci.*, vol. 1, pp. 329–334, Mar 1992.

[80] J. Jeffery, B. Persson, I. Wood, T. Bergman, R. Jeffery, and H. Jornvall, "Glucose-6-phosphate dehydrogenase. Structure-function relationships and the Pichia jadinii enzyme structure," *Eur. J. Biochem.*, vol. 212, pp. 41–49, Feb 1993.

[81] M. Kotaka, S. Gover, L. Vandeputte-Rutten, S. W. Au, V. M. Lam, and M. J. Adams, "Structural studies of glucose-6-phosphate and NADP+ binding to human glucose-6-phosphate dehydrogenase," *Acta Crystallogr. D Biol. Crystallogr.*, vol. 61, pp. 495–504, May 2005.

[82] J. M. Bautista, P. J. Mason, and L. Luzzatto, "Human glucose-6-phosphate dehydrogenase. Lysine 205 is dispensable for substrate binding but essential for catalysis," *FEBS Lett.*, vol. 366, pp. 61–64, Jun 1995.

[83] M. Born and R. Oppenheimer, "Zur quantentheorie der molekeln," *Ann. Phys. (Leipzig)*, vol. 84 (20), p. 457, 1927.

[84] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "Charmm: A program for macromolecular energy, minimization,

and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, no. 2, pp. 187–217, 1983.

[85] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, and J. W. Caldwell, "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules," *Journal of The American Chemical Society*, vol. 117, pp. 5179–5197, 1995.

[86] M. Christen, P. H. Hunenberger, D. Bakowies, R. Baron, R. Burgi, D. P. Geerke, T. N. Heinz, M. A. Kastenholz, V. Krautler, C. Oostenbrink, C. Peter, D. Trzesniak, and W. F. van Gunsteren, "The GROMOS software for biomolecular simulation: GROMOS05," *J Comput Chem*, vol. 26, pp. 1719–1751, Dec 2005.

[87] A. D. Mackerell, "Empirical force fields for biological macromolecules: overview and issues," *J Comput Chem*, vol. 25, pp. 1584–1604, Oct 2004.

[88] C. J. Cramer, *Essentials of Computational Chemistry Theories and Models - second edition*. WILEY, 2004.

[89] E. W. Weisstein, ""newton's method."," *From MathWorld–A Wolfram Web Resource.*

[90] C. J. Kwok, A. C. Martin, S. W. Au, and V. M. Lam, "G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations," *Hum. Mutat.*, vol. 19, pp. 217–224, Mar 2002.

[91] J. M. Hurst, L. E. McMillan, C. T. Porter, J. Allen, A. Fakorede, and A. C. Martin, "The SAAPdb web resource: a large-scale structural analysis of mutant proteins," *Hum. Mutat.*, vol. 30, pp. 616–624, Apr 2009.

[92] A. C. Martin, A. M. Facchiano, A. L. Cuff, T. Hernandez-Boussard, M. Olivier, P. Hainaut, and J. M. Thornton, "Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein," *Hum. Mutat.*, vol. 19, pp. 149–164, Feb 2002.

[93] N. S. Al-Numail and A. C. Martin, "The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations," *BMC Genomics*, vol. 14 Suppl 3, p. S4, 2013.

[94] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, pp. 3812–3814, Jul 2003.

[95] P. D. Thomas, M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania, "PANTHER: a library of protein families and subfamilies indexed by function," *Genome Res.*, vol. 13, pp. 2129–2141, Sep 2003.

[96] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: application to cancer genomics," *Nucleic Acids Res.*, vol. 39, p. e118, Sep 2011.

[97] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nat. Methods*, vol. 7, pp. 248–249, Apr 2010.

[98] A. Gonzalez-Perez and N. Lopez-Bigas, "Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel," *Am. J. Hum. Genet.*, vol. 88, pp. 440–449, Apr 2011.

[99] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[100] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update."

[101] A. Minucci, K. Moradkhani, M. J. Hwang, C. Zuppi, B. Giardina, and E. Capoluongo, "Glucose-6-phosphate dehydrogenase (G6PD) mutations database: review of the "old" and update of the new mutations," *Blood Cells Mol. Dis.*, vol. 48, pp. 154–165, Mar 2012.

[102] A. C. Martin, "Mutmodel v1.17," *UCL*, 1996-2011.

[103] H. H. Shih, J. Brady, and M. Karplus, "Structure of proteins with single-site mutations: a minimum perturbation approach," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 82, pp. 1697–1700, Mar 1985.

[104] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "Gromacs 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," *Bioinformatics*, vol. 29, no. 7, pp. 845–854, 2013.

[105] H. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs: A message-passing parallel molecular dynamics implementation," *Computer Physics Communications*, pp. 43 – 56, 1995.

[106] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "Gromacs: Fast, flexible, and free," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005.

[107] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.

[108] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field," *Proteins*, vol. 78, pp. 1950–1958, Jun 2010.

[109] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of Chemical Physics*, vol. 79, no. 2, pp. 926–935, 1983.

[110] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: The gromos force-field parameter sets 53a5 and 53a6," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.

[111] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic validation of protein force fields against experimental data," *PLoS ONE*, vol. 7, no. 2, p. e32131, 2012.

[112] M. Rueda, C. Ferrer-Costa, T. Meyer, A. Perez, J. Camps, A. Hospital, J. L. Gelpi, and M. Orozco, "A consensus view of protein dynamics," *Proceedings of the National Academy of Sciences*, vol. 104, no. 3, pp. 796–801, 2007.

[113] T. Darden, D. York, and L. Pedersen, "Particle mesh ewald: An nlog(n) method for ewald sums in large systems," *The Journal of Chemical Physics*, vol. 98, no. 12, 1993.

[114] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of Chemical Physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[115] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," *Journal of Applied Physics*, vol. 52, no. 12, pp. 7182–7190, 1981.

[116] W. G. Hoover, "Canonical dynamics: Equilibrium phase-space distributions," *Phys. Rev. A*, vol. 31, pp. 1695–1697, Mar 1985.

[117] S. Nosé, "A unified formulation of the constant temperature molecular dynamics methods," *The Journal of Chemical Physics*, vol. 81, no. 1, pp. 511–519, 1984.

[118] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije, "Lincs: A linear constraint solver for molecular simulations," *Journal of Computational Chemistry*, vol. 18, no. 12, pp. 1463–1472, 1997.

[119] J.-P. Ryckaert, G. Ciccotti, and H. J. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes," *Journal of Computational Physics*, vol. 23, no. 3, pp. 327 – 341, 1977.

[120] W. B. Langdon, "Initial experiences of the emerald: e-infrastructure south GPU supercomputer," Research Note RN/12/08, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK, 17 June 2012.

[121] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[122] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer New York, 2009.

[123] R. J. Pantazes and C. D. Maranas, "OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding," *Protein Eng. Des. Sel.*, vol. 23, pp. 849–858, Nov 2010.

[124] *Molecular modeling of proteins 1st edition*. Andreas Kukol.

[125] K. Suhre and Y.-H. Sanejouand, "Elnémo: a normal mode web server for protein movement analysis and the generation of templates for molecular replacement," *Nucleic Acids Research*, vol. 32, no. suppl 2, pp. W610–W614, 2004.

[126] K. Suhre and Y.-H. Sanejouand, "On the potential of normal-mode analysis for solving difficult molecular-replacement problems," *Acta Crystallographica Section D*, vol. 60, pp. 796–799, Apr 2004.

[127] B. Halle, "Flexibility and packing in proteins," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 1274–1279, Feb. 2002.

[128] D. Roos, R. van Zwieten, J. T. Wijnen, F. Gomez-Gallego, M. de Boer, D. Stevens, C. J. Pronk-Admiraal, T. de Rijk, C. J. van Noorden, R. S. Weening, T. J. Vulliamy, J. E. Ploem, P. J. Mason, J. M. Bautista, P. M. Khan, and E. Beutler, "Molecular basis and enzymatic properties of glucose 6-phosphate dehydrogenase volendam, leading to chronic nonspherocytic anemia, granulocyte dysfunction, and increased susceptibility to infections," *Blood*, vol. 94, pp. 2955–2962, Nov 1999.

[129] C. Neale, R. Pomes, and A. E. Garcia, "Peptide Bond Isomerization in High-Temperature Simulations," *J Chem Theory Comput*, Mar 2016.

[130] L. T. Chao, C. S. Du, E. Louie, L. Zuo, E. Chen, B. Lubin, and D. T. Chiu, "A to G substitution identified in exon 2 of the G6PD gene among G6PD deficient Chinese," *Nucleic Acids Res.*, vol. 19, p. 6056, Nov 1991.

[131] T. J. Vulliamy, M. D'Urso, G. Battistuzzi, M. Estrada, N. S. Foulkes, G. Martini, V. Calabro, V. Poggi, R. Giordano, and M. Town, "Diverse point mutations in the human glucose-6-phosphate dehydrogenase gene cause enzyme deficiency and mild or severe hemolytic anemia," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 85, pp. 5171–5175, Jul 1988.

[132] O. Babalola, R. Cancedda, and L. Luzzatto, "Genetic variants of glucose 6-phosphate dehydrogenase from human erythrocytes: unique properties of the A - variant isolated from "deficient" cells," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 69, pp. 946–950, Apr 1972.

[133] F. Gomez-Gallego, A. Garrido-Pertierra, and J. M. Bautista, "Structural defects underlying protein dysfunction in human glucose-6-phosphate dehydrogenase A(-) deficiency," *J. Biol. Chem.*, vol. 275, pp. 9256–9262, Mar 2000.

[134] M. Ganczakowski, M. Town, D. K. Bowden, T. J. Vulliamy, A. Kaneko, J. B. Clegg, D. J. Weatherall, and L. Luzzatto, "Multiple glucose 6-phosphate dehydrogenase-deficient variants correlate with malaria endemicity in the Vanuatu archipelago (southwestern Pacific)," *Am. J. Hum. Genet.*, vol. 56, pp. 294–301, Jan 1995.

[135] R. Zarza, A. Pujades, A. Rovira, R. Saavedra, J. Fernandez, M. Aymerich, and J. L. Vives Corrons, "Two new mutations of the glucose-6-phosphate dehydrogenase (G6PD) gene associated with haemolytic anaemia: clinical, biochemical and molecular relationships," *Br. J. Haematol.*, vol. 98, pp. 578–582, Sep 1997.

[136] X. Ren, Y. He, C. Du, W. Jiang, L. Chen, and Q. Lin, "A novel mis-sense mutation (G1381A) in the G6PD gene identified in a Chinese man," *Chin. Med. J.*, vol. 114, pp. 399–401, Apr 2001.

[137] E. Beutler, B. Westwood, J. T. Prchal, G. Vaca, C. S. Bartsocas, and L. Baronciani, "New glucose-6-phosphate dehydrogenase mutations from various ethnic groups," *Blood*, vol. 80, pp. 255–256, Jul 1992.

[138] T. Takizawa, H. Fujii, S. Takegawa, K. Takahashi, A. Hirono, T. Morisaki, H. Kanno, R. Oka, H. Yoshioka, and S. Miwa, "A unique electrophoretic slow-moving glucose 6-phosphate dehydrogenase variant (G6PD Asahikawa) with a markedly acidic pH optimum," *Hum. Genet.*, vol. 68, no. 1, pp. 70–72, 1984.

[139] I. Sillitoe, T. E. Lewis, A. Cuff, S. Das, P. Ashford, N. L. Dawson, N. Furnham, R. A. Laskowski, D. Lee, J. G. Lees, *et al.*, "CATH: comprehensive structural and functional annotations for genome sequences," *Nucleic acids research*, p. gku947, 2014.

[140] I. G. Kevrekidis, C. W. Gear, and G. Hummer, "Equation-free: The computer-aided analysis of complex multiscale systems," *AIChE Journal*, vol. 50, no. 7, pp. 1346–1355, 2004.

[141] C. Theodoropoulos, Y. H. Qian, and I. G. Kevrekidis, ""Coarse" stability and bifurcation analysis using time-steppers: a reaction-diffusion example," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 97, pp. 9840–9843, Aug 2000.

[142] A. Laio and F. L. Gervasio, "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," *Reports on Progress in Physics*, vol. 71, no. 12, p. 126601, 2008.

[143] D. Cvijović and J. Klinowski, *Handbook of Global Optimization: Volume 2*, ch. Taboo Search: An Approach to the Multiple-Minima Problem for Continuous Functions, pp. 387–406. Boston, MA: Springer US, 2002.

[144] T. Huber, A. E. Torda, and W. F. van Gunsteren, "Local elevation: A method for improving the searching properties of molecular dynamics simulation," *Journal of Computer-Aided Molecular Design*, vol. 8, no. 6, pp. 695–708, 1994.

[145] E. Zurek and W. Grochala, "Predicting crystal structures and properties of matter under extreme conditions via quantum mechanics: the pressure is on," *Phys. Chem. Chem. Phys.*, vol. 17, pp. 2917–2934, 2015.

[146] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello, "Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics," *J. Am. Chem. Soc.*, vol. 128, pp. 13435–13441, Oct 2006.

[147] K. Gaalswyk and C. N. Rowley, "An explicit-solvent conformation search method using open software," *PeerJ*, vol. 4, p. e2088, 2016.

[148] M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, F. Marinelli, F. Pietrucci, R. A. Broglia, and M. Parrinello, "Plumed: A portable plugin for free-energy calculations with molecular dynamics," *Computer Physics Communications*, vol. 180, no. 10, pp. 1961 – 1972, 2009.

[149] P. Rowland, A. K. Basak, S. Gover, H. R. Levy, and M. J. Adams, "The three-dimensional structure of glucose 6-phosphate dehydrogenase from Leuconostoc mesenteroides refined at 2.0 A resolution," *Structure*, vol. 2, pp. 1073–1087, Nov 1994.

[150] G. G. Maisuradze, P. Senet, C. Czaplewski, A. Liwo, and H. A. Scheraga, "Investigation of protein folding by coarse-grained molecular dynamics with the UNRES force field," *J Phys Chem A*, vol. 114, pp. 4471–4485, Apr 2010.

[151] A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Odziej, K. Wachucik, and H. A. Scheraga, "Modification and optimization of the united-residue (UNRES) potential energy function for canonical simulations. I. Temperature dependence of the effective energy function and tests of the optimization method with single training proteins," *J Phys Chem B*, vol. 111, pp. 260–285, Jan 2007.

[152] U. Kozowska, G. G. Maisuradze, A. Liwo, and H. A. Scheraga, "Determination of side-chain-rotamer and side-chain and backbone virtual-bond-stretching potentials of mean force from AM1 energy surfaces of terminally-blocked amino-acid residues, for coarse-grained simulations of protein structure and folding. II. Results, comparison with statistical potentials, and implementation in the UNRES force field," *J Comput Chem*, vol. 31, pp. 1154–1167, Apr 2010.

[153] Y. He, Y. Xiao, A. Liwo, and H. A. Scheraga, "Exploring the parameter space of the coarse-grained UNRES force field by random search: selecting a transferable medium-resolution force field," *J Comput Chem*, vol. 30, pp. 2127–2135, Oct 2009.

[154] U. Kozowska, A. Liwo, and H. A. Scheraga, "Determination of side-chain-rotamer and side-chain and backbone virtual-bond-stretching potentials of mean force from AM1 energy surfaces of terminally-blocked amino-acid residues, for coarse-grained simulations of protein structure and folding. I. The method," *J Comput Chem*, vol. 31, pp. 1143–1153, Apr 2010.

[155] J. G. Gay and B. J. Berne, "Modification of the overlap potential to mimic a linear site-site potential," *The Journal of Chemical Physics*, vol. 74, no. 6, 1981.

[156] J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, and H. A. Scheraga, "Calculation of protein conformation by global optimization of a potential energy function," *Proteins*, vol. Suppl 3, pp. 204–208, 1999.

[157] J. Pillardy, C. Czaplewski, A. Liwo, J. Lee, D. R. Ripoll, R. Ka?mierkiewicz, S. Oldziej, W. J. Wedemeyer, K. D. Gibson, Y. A. Arnautova, J. Saunders, Y. J. Ye, and H. A. Scheraga, "Recent improvements in prediction of protein structure by global optimization of a potential energy function," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 98, pp. 2329–2333, Feb 2001.

[158] M. Khalili, A. Liwo, F. Rakowski, P. Grochowski, and H. A. Scheraga, "Molecular dynamics with the united-residue model of polypeptide chains. I. Lagrange equations of motion and tests of numerical stability in the microcanonical mode," *J Phys Chem B*, vol. 109, pp. 13785–13797, Jul 2005.

[159] A. Liwo, M. Khalili, and H. A. Scheraga, "Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 2362–2367, Feb 2005.

[160] A. V. Rojas, A. Liwo, and H. A. Scheraga, "Molecular dynamics with the United-residue force field: ab initio folding simulations of multichain proteins," *J Phys Chem B*, vol. 111, pp. 293–309, Jan 2007.

[161] G. G. Maisuradze, A. Liwo, and H. A. Scheraga, "Principal component analysis for protein folding dynamics," *J. Mol. Biol.*, vol. 385, pp. 312–329, Jan 2009.

[162] D. D. Humphreys, R. A. Friesner, and B. J. Berne, "A multiple-time-step molecular dynamics algorithm for macromolecules," *The Journal of Physical Chemistry*, vol. 98, no. 27, pp. 6885–6892, 1994.

[163] M. Levitt and J. Greer, "Automatic identification of secondary structure in globular proteins," *J. Mol. Biol.*, vol. 114, pp. 181–239, Aug 1977.

[164] J. H. W. Jr., "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[165] W. Yan, J. Zhou, M. Sun, J. Chen, G. Hu, and B. Shen, "The construction of an amino acid network for understanding protein structure and function," *Amino Acids*, vol. 46, pp. 1419–1439, Jun 2014.

[166] S. Khor, "Towards an integrated understanding of the structural characteristics of protein residue networks," *Theory in Biosciences*, vol. 131, no. 2, pp. 61–75, 2012.

[167] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, "Small-world view of the amino acids that play a key role in protein folding," *Phys. Rev. E*, vol. 65, Jun 2002.

[168] L. D. Paola, M. D. Ruvo, P. Paci, D. Santoni, and A. Giuliani, "Protein contact networks: An emerging paradigm in chemistry," *Chemical Reviews*, vol. 113, no. 3, pp. 1598–1613, 2013.

[169] P. Csermely, "Creative elements: network-based predictions of active centres in proteins and cellular and social networks," *Trends in Biochemical Sciences*, vol. 33, no. 12, pp. 569 – 576, 2008.

[170] D. Piovesan, G. Minervini, and S. C. Tosatto, "The ring 2.0 web server for high quality residue interaction networks," 2016.

[171] N. T. Doncheva, K. Klein, F. S. Domingues, and M. Albrecht, "Analyzing and visualizing residue networks of protein structures," *Trends in Biochemical Sciences*, vol. 36, no. 4, pp. 179 – 182, 2011.

[172] M. Munz and P. C. Biggin, "JGromacs: a Java package for analyzing protein simulations," *J Chem Inf Model*, vol. 52, pp. 255–259, Jan 2012.

[173] J. Eargle and Z. Luthey-Schulten, "NetworkView: 3D display and analysis of proteinÂůRNA interaction networks," *Bioinformatics*, vol. 28, pp. 3000–3001, Nov 2012.

[174] M. Bhattacharyya, C. R. Bhat, and S. Vishveshwara, "An automated approach to network features of protein structure ensembles," *Protein Sci.*, vol. 22, pp. 1399–1416, Oct 2013.

[175] M. Pasi, M. Tiberti, A. Arrigoni, and E. Papaleo, "xPyder: a PyMOL plugin to analyze coupled residues and their networks in protein structures," *J Chem Inf Model*, vol. 52, pp. 1865–1874, Jul 2012.

[176] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus, "Three key residues form a critical contact network in a protein folding transition state," *Nature*, vol. 409, pp. 641–645, Feb 2001.

[177] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, "Small-world view of the amino acids that play a key role in protein folding," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 65, p. 061910, Jun 2002.

[178] G. Bagler and S. Sinha, "Assortative mixing in Protein Contact Networks and protein folding kinetics," *Bioinformatics*, vol. 23, pp. 1760–1767, Jul 2007.

[179] Y. Li, G. Li, Z. Wen, H. Yin, M. Hu, J. Xiao, and M. Li, "Novel feature for catalytic protein residues reflecting interactions with other residues," *PLoS ONE*, vol. 6, pp. 1–9, 03 2011.

[180] K. V. Brinda and S. Vishveshwara, "A network representation of protein structures: implications for protein stability," *Biophys. J.*, vol. 89, pp. 4159–4170, Dec 2005.

[181] X. Jiao and S. Chang, "Scoring function based on weighted residue network," *Int J Mol Sci*, vol. 12, no. 12, pp. 8773–8786, 2011.

[182] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski, "Network analysis of protein structures identifies functional residues," *J. Mol. Biol.*, vol. 344, pp. 1135–1146, Dec 2004.

[183] K. Z. Szalay and P. Csermely, "Perturbation centrality and turbine: A novel centrality measure obtained using a versatile network dynamics tool," *PLoS ONE*, vol. 8, 10 2013.

[184] M. Seeber, A. Felline, F. Raimondi, S. Muff, R. Friedman, F. Rao, A. Caflisch, and F. Fanelli, "Wordom: a user-friendly program for the analysis of molecular structures, trajectories, and free energy surfaces," *J Comput Chem*, vol. 32, pp. 1183–1194, Apr 2011.

[185] M. Seeber, M. Cecchini, F. Rao, G. Settanni, and A. Caflisch, "Wordom: a program for efficient analysis of molecular dynamics simulations," *Bioinformatics*, vol. 23, pp. 2625–2627, Oct 2007.

[186] J. A. McCammon, "Protein dynamics," *Reports on Progress in Physics*, vol. 47, no. 1, p. 1, 1984.

[187] N. Kannan and S. Vishveshwara, "Identification of side-chain clusters in protein structures by a graph spectral method," *J. Mol. Biol.*, vol. 292, pp. 441–464, Sep 1999.

[188] D. J. Watts and S. H. Strogatz, "Collective dynamics of'small-world'networks.," *Nature*, vol. 393, no. 6684, pp. 409–10, 1998.

[189] R. P. Joosten, T. A. te Beek, E. Krieger, M. L. Hekkelman, R. W. Hooft, R. Schneider, C. Sander, and G. Vriend, "A series of PDB related databases for everyday needs," *Nucleic Acids Res.*, vol. 39, pp. D411–419, Jan 2011.